Aalto University
School of Science

Lauri Mustonen

Parametric differential equations and inverse diffusivity problem

Master's thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology in the Degree Programme in Engineering Physics and Mathematics.

Espoo, 27th May 2014

Supervisor: Associate Professor Nuutti Hyvönen
Instructor: M.Sc. (Tech.) Matti Leinonen

Aalto University
School of Science

Aalto University School of Science	Abstract of the Master's Thesis

Author: Lauri Mustonen

Title: Parametric differential equations and inverse diffusivity problem

Degree Programme: Degree Programme in Engineering Physics and Mathematics

Major subject: Mathematics
Minor subject: Information and Computer Science

Chair (code): Mat-1

Supervisor: Associate Professor Nuutti Hyvönen
Instructor: M.Sc. (Tech.) Matti Leinonen

Abstract: Parametric differential equations have received increased attention during the past decade, mostly due to their applications to quantifying stochastic systems. In this thesis, we formulate the parametric time-dependent diffusion equation and solve it numerically by using the finite element method in the spatial domain and the spectral Galerkin method in the parameter domain. The obtained solution is used as a tool for an inverse boundary value problem, where the unknown is the diffusion coefficient. The absence of random variables in the forward problem allows choosing compactly supported functions such as splines to represent the diffusivity, whereas the stochastic equations are usually parametrized by using orthogonal Karhunen–Loève eigenfunctions. We analyze how the locality of the diffusivity functions affects the sparsity of the resulting large linear system. In the context of direct equation solvers, the reduction of fill-in is addressed with numerical examples. Although discretization errors are clearly visible, diffusivity reconstructions indicate that the solution to the parametric equation may provide a feasible algorithm for the inverse diffusivity problem, which further can be considered as a basis for thermal tomography.

Date: 27th May 2014	**Language:** English	**Number of pages:** v+54

Keywords: parametric differential equations, inverse problems, sparse matrices, splines

Aalto-yliopisto
Perustieteiden korkeakoulu

Aalto-yliopisto Perustieteiden korkeakoulu	Diplomityön tiivistelmä
Tekijä: Lauri Mustonen	
Työn nimi: Parametriset differentiaaliyhtälöt ja diffuusiokertoimen käänteisongelma **Työn nimi (engl.):** Parametric differential equations and inverse diffusivity problem	
Tutkinto-ohjelma: Teknillisen fysiikan ja matematiikan tutkinto-ohjelma	
Pääaine: Matematiikka **Sivuaine:** Tietojenkäsittelytiede	
Opetusyksikön (ent. professuuri) koodi: Mat-1	
Työn valvoja: Professori Nuutti Hyvönen **Työn ohjaaja:** DI Matti Leinonen	

Tiivistelmä: Kiinnostus parametririippuvia differentiaaliyhtälöitä kohtaan on kasvanut viimeisen vuosikymmenen aikana. Pääosin tämä on johtunut satunnaisuutta sisältäviin malleihin liittyvistä sovelluskohteista. Tässä työssä esitellään parametrinen ajasta riippuva diffuusioyhtälö, joka ratkaistaan numeerisesti käyttämällä elementtimenetelmää paikan suhteen ja Galerkinin spektraalimenetelmää parametrialueessa. Saatua ratkaisua hyödynnetään käänteisessä reuna-arvo-ongelmassa, jossa tuntemattomana suureena on diffuusiokerroin. Satunnaismuuttujien puuttuminen suorassa ongelmassa mahdollistaa diffuusiokertoimen esittämisen kompaktikantajaisten funktioiden avulla. Näin voidaan käyttää esimerkiksi palapolynomeja eli splinejä, kun taas stokastiset yhtälöt parametrisoidaan yleensä käyttämällä ortogonaalisia Karhunen–Loève-ominaisfunktioita. Työssä analysoidaan lokaalien funktioiden vaikutusta lineaarisen yhtälöryhmän harvuuteen. Lisäksi matriisien täyttymisen vähentymistä suorien yhtälöratkaisimien yhteydessä tutkitaan numeerisin esimerkein. Vaikka diskretointivirheet näkyvät selvästi, diffusiokertoimesta muodostettujen rekonstruktioiden perusteella näyttää siltä, että parametrisen yhtälön avulla saadaan toimiva ratkaisualgoritmi diffuusiokertoimen käänteisongelmaan, jota puolestaan voidaan pitää lämpötomografian perustana.

Päivämäärä: 27.5.2014	**Kieli:** englanti	**Sivumäärä:** v+54
Avainsanat: parametriset differentiaaliyhtälöt, inversio-ongelmat, harvat matriisit, splinit		

Preface

This master's thesis was written as a member of the inverse problems research group in Aalto University. As a result of the creative atmosphere established by my supervisor Nuutti Hyvönen, it was possible to adapt the work during the writing process according to the new ideas stemming from this rapidly developing field of mathematics. While being under guidance of Hyvönen for two years, this thesis is also largely influenced by the research done by my instructor Matti Leinonen.

In addition to my supervisor and instructor, I would also like to thank other students and staff with whom I have been collaborating during the past few years. Especially I would like to thank all those awesome people who I met in Denmark last autumn.

Espoo, 27th May 2014

Lauri Mustonen

Contents

1 Introduction **1**

2 Parametric diffusion equation **4**
2.1 Model problem 4
2.2 Semi-discretized equation 6
2.3 Spatial basis functions 8
2.4 Basis functions for parameter domain 12

3 Computations **16**
3.1 Constructing the system 16
3.2 Sparsity 19
3.3 Solving the system 25

4 Inverse diffusivity problem **30**
4.1 Estimating parameters from partial data 31
4.2 Regularization and Bayesian inversion 33

5 Numerical examples **37**
5.1 Convergence of parametric solution 37
5.2 Inverse problem and reconstructions 42

6 Conclusions and future work **50**

References **52**

Chapter 1

Introduction

Partial differential equations arising from applications usually involve an unknown function of only few variables. Often, these variables represent the spatial coordinates and possibly time. A numerical or analytical solution to the equation consists of a function that in some sense approximates or coincides with the unknown (say, primary) function. When the unknown function depends not only on spatial and temporal variables, but also on another (secondary) function, the number of variables in the problem and its solution may become very large or even infinite. Such a situation occurs, for example, when the equation involves a material property that functionally varies within an object and the solution is sought in the form that explicitly maps a given property function, accompanied by physical coordinates, to a real number. In that case, one can think of having a family of differential operators, each corresponding to one particular secondary function, and a single solution that contains the information of the whole operator family. The secondary function could also be related, for example, to initial or boundary conditions or to the geometry of the problem. In practice, parametrizing the secondary function by a moderate number of variables is necessary in order to construct a computationally tractable problem. The resulting equation is called *parametric (partial) differential equation*. Typically, it does not contain derivatives of the parameter variables.

There are several reasons why considering a parametric equation is of interest. First, if the secondary function is a stochastic process or a random field, the solution is also a random quantity and one may want to obtain statistics such as mean and variance of the solution. The solution statistics can often be computed efficiently, provided that the randomness is propagated to the solution by parametrizing the secondary function by random variables and then expressing the solution as a function of those same random variables. This kind of *uncertainty quantification* has quite recently

become an attractive alternative to traditional methods like Monte Carlo sampling, where the solution statistics are computed after solving a large set of deterministic, non-parametric problems, where in each problem the secondary function is drawn randomly according to its probability distribution. The stochastic framework has influenced both terminology and methods related to parametric equations, which in that context are usually called just stochastic equations. Books such as [17] and [22] provide detailed yet practical material for stochastic computations and they also serve as good general references for the subjects considered in this thesis.

Another motivation for parametric differential equations stems from a class of *inverse problems*, where the secondary function is to be estimated based on some observations about the primary function. The most straightforward way to approach such an inverse problem is to iteratively solve equations resulting from different secondary functions until a sufficient match between the observations and the computed values is obtained. Solving the problem once for all secondary functions can reduce the computational cost of the iterative stage by orders of magnitude, which is especially advantageous if the inverse problem is solved several times for different measurement data. In chapter 4, we will investigate inverse problems as applications of parametric differential equations in more detail.

This thesis is organized as follows. Chapter 2 formulates and discretizes the parametric diffusion equation that will be used as a model problem in the rest of the thesis. In particular, we discuss how to parametrize the diffusion coefficient and how to discretize the spatial and parametric function spaces where the solution is defined. In chapter 3, we treat the computational issues that arise from the large dimension of the discrete linear system. Matrix construction, sparsity and fill-in are considered in detail. The discretization of the temporal domain is also discussed. As mentioned above, chapter 4 addresses an inverse problem, more precisely the inverse diffusivity problem, where the objective is to determine the diffusion coefficient based on boundary measurements. Classical regularization techniques and their necessity for inverse problems are quickly reviewed. We also touch the stochastic framework from the Bayesian point of view by interpreting the diffusivity, measurements and the outcome of the inverse problem as random variables. Although chapter 3 provides some numerical examples as well, the goal in chapter 5 is to thoroughly illustrate the developed methods with chosen test cases. We first show how the numerical approximation to the parametric equation behaves compared to the exact solution, after which we present a few diffusivity reconstructions related to the inverse diffusivity problem. Finally, chapter 6 concludes this work and suggests some possible research topics for the future.

The approach of this thesis is computational. Instead of theoretical results, we put emphasis on practical issues that arise during numerical implementation. Splines have not been widely used to parametrize the coefficient function in a parametric differential equation, and the remarks in chapter 3, concerning the sparsity that follows from compactly supported splines, can be considered as new results. Furthermore, inverse boundary value problems for parabolic equations have not received very much attention in the literature. Especially, diffusivity reconstructions based on parametric solution, such as those presented in section 5.2, are unlikely to be found in previous research papers.

Chapter 2

Parametric diffusion equation

The diffusion equation can be used to describe many physical phenomena such as heat transfer in a media. In this chapter, we define the parametric diffusion equation and explain how it can be discretized in both spatial and parametric dimensions. The reader is assumed to be more or less familiar with partial differential equations and Galerkin methods. For an introductory treatment of those topics, we recommend [9]. Some basic concepts related to splines and orthogonal polynomials are quickly surveyed in sections 2.3 and 2.4, respectively.

2.1 Model problem

Before turning the discussion to a parametric partial differential equation, let us first consider the corresponding non-parametric problem. Throughout this thesis, we will assume that the spatial domain $\Omega \subset \mathbf{R}^d$, where $d \in \{1, 2, 3\}$, is bounded and has a Lipschitz boundary $\Gamma \coloneqq \partial\Omega$ which is piecewise smooth such that the outer-pointing unit normal vector $\hat{\boldsymbol{n}}$ is well-defined almost everywhere. For the sake of concreteness, we will restrict our attention to the parabolic initial value problem

$$\begin{aligned} \partial_t u(\boldsymbol{x}, t) - \nabla \cdot (a(\boldsymbol{x}) \nabla u(\boldsymbol{x}, t)) &= f(\boldsymbol{x}, t), \qquad \boldsymbol{x} \in \Omega, \quad 0 < t \leq T, \\ u(\boldsymbol{x}, 0) &= u_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \end{aligned} \tag{2.1}$$

where $T > 0$ is the final time and all functions are real-valued. This can be thought of as the strong form of an unsteady diffusion equation with a describing the isotropic time-independent diffusivity. The problem is equipped with mixed Dirichlet–Neumann boundary conditions

$$\begin{aligned} u(\boldsymbol{x}, t) &= 0, \qquad \boldsymbol{x} \in \Gamma_{\mathrm{D}}, \\ a(\boldsymbol{x}) \nabla u(\boldsymbol{x}, t) \cdot \hat{\boldsymbol{n}}(\boldsymbol{x}) &= g(\boldsymbol{x}, t), \qquad \boldsymbol{x} \in \Gamma_{\mathrm{N}}, \end{aligned} \tag{2.2}$$

where $\Gamma_\mathrm{D} \cap \Gamma_\mathrm{N} = \emptyset$, $\Gamma = \overline{\Gamma_\mathrm{D}} \cup \overline{\Gamma_\mathrm{N}}$ and one of the two parts can be empty. The interpretation of the derivatives and the boundaries in the case $d = 1$ should be obvious. We assume $u_0 \in L^2(\Omega)$ and $a \in L^\infty(\Omega)$ such that

$$a_\mathrm{min} \leq a(\boldsymbol{x}) \leq a_\mathrm{max} \tag{2.3}$$

almost everywhere in Ω for some constants $0 < a_\mathrm{min}, a_\mathrm{max} < \infty$. The source (or forcing) term f and the Neumann boundary value g are assumed to be Lipschitz continuous with respect to time t. For any fixed time, we further assume that $f(t) \in L^2(\Omega)$ and $g(t) \in L^2(\Gamma_\mathrm{N})$.

We re-formulate the problem in a weak sense. To this end, both sides of the equation (2.1) are multiplied by a time-independent test function v and integrated over the spatial domain. After performing integration by parts, we obtain

$$\begin{aligned} \partial_t (u, v)_{L^2(\Omega)} + (a \nabla u, \nabla v)_{L^2(\Omega)} &= (f, v)_{L^2(\Omega)} + (g, v)_{L^2(\Gamma_\mathrm{N})}, \\ (u, v)_{L^2(\Omega)} &= (u_0, v)_{L^2(\Omega)}, \end{aligned} \tag{2.4}$$

where $(\cdot, \cdot)$ denotes the standard L^2 inner product and the latter equation is valid for $t = 0$. If the equalities are required to hold for all test functions from the space

$$V \coloneqq \{v \in H^1(\Omega) \mid v_{|\Gamma_\mathrm{D}} = 0\}, \tag{2.5}$$

where H^1 denotes the first-order Sobolev space (i.e., functions in $L^2(\Omega)$ whose first-order weak derivatives are square-integrable [9, Chap. 5]) and the restriction to the boundary is understood in the sense of traces, the problem (2.4) admits a unique solution (in the sense of L^2 equivalence classes) $u(t) \in V$ for each $t \in [0, T]$.

The problem (2.4) may admit a (unique) steady-state solution, corresponding to $\partial_t u = 0$, if f and g have limits with respect to t. If $\Gamma_\mathrm{D} \neq \emptyset$, this is a sufficient condition. Otherwise, we also require the compatibility condition

$$(f, 1)_{L^2(\Omega)} + (g, 1)_{L^2(\Gamma)} = 0$$

for the limit functions of f and g to be satisfied in order to have a steady-state solution. If we are only interested in the steady state, the problem reduces to an elliptic equation. Indeed, most of the methods developed in this chapter apply to a broader class of problems than the one defined by equations (2.1) and (2.2).

The diffusivity a can be considered as a secondary function in a parametric partial differential equation. To that end, let $a = a(\boldsymbol{x}, \boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta} \in \Theta \subseteq \mathbf{R}^P$ is a vector of parameter variables such that for each fixed $\boldsymbol{\vartheta}$ the diffusivity function a belongs to $L^\infty(\Omega)$ and satisfies the condition (2.3). Consequently,

the solution also depends on this parameter vector, that is, $u = u(\boldsymbol{x}, t, \boldsymbol{\vartheta})$. Taking into account the assumptions, we see that for every fixed $\boldsymbol{\vartheta} \in \Theta$ the function $u(\boldsymbol{x}, t, \boldsymbol{\vartheta})$ is the unique solution to the problem (2.4). Thus, the unique solvability of the parametric equation follows immediately. Let us still re-formulate the problem in a weak sense. The weak formulation is now derived from (2.1) after multiplying the equation by a test function and integrating over the domain $\Omega \times \Theta$. This results in an initial value problem

$$\begin{aligned} \partial_t(u, v)_{L^2_w(\Omega\times\Theta)} + (a\nabla u, \nabla v)_{L^2_w(\Omega\times\Theta)} &= (f, v)_{L^2_w(\Omega\times\Theta)} + (g, v)_{L^2_w(\Gamma_{\mathrm{N}}\times\Theta)}, \\ (u, v)_{L^2_w(\Omega\times\Theta)} &= (u_0, v)_{L^2_w(\Omega\times\Theta)}, \end{aligned} \tag{2.6}$$

where $u(t), v \in V \otimes L^2_w(\Theta)$ and the second equality is again valid for $t = 0$. Here, $\otimes$ denotes a tensor product of Hilbert spaces (see e.g. [20]). Naturally, the derivatives in (2.6) affect only the spatial variables. The inner products in (2.6) are performed in a weighted L^2 space for reasons that will become clear shortly. More precisely, we define

$$(u, v)_{L^2_w(D\times\Theta)} \coloneqq \int_\Theta \int_D u(\boldsymbol{x}, \boldsymbol{\vartheta}) v(\boldsymbol{x}, \boldsymbol{\vartheta}) w(\boldsymbol{\vartheta}) \,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\vartheta}$$

for a positive weight function $w\colon \Theta \to \mathbf{R}_+$. Here, the spatial domain D can be either Ω or Γ_{N}.

2.2 Semi-discretized equation

Next, we will discretize the model problem (2.6). As suggested by the integral formulation, we will use the Galerkin method for spatial and parametric dimensions and later handle the temporal dimension separately. This is not the only possibility, since the lack of derivatives with respect to the parameter variables would allow a simple interpolatory approach where the problem (2.4) is solved for several fixed values of $\boldsymbol{\vartheta}$ by using standard tools. The solution $u(\boldsymbol{x}, t, \boldsymbol{\vartheta})$ would then be written by using interpolation rules for the parametric dimension. The drawback of this collocative or *non-intrusive* approach is the large number of problems that have to be solved. This is in particular the case when the dimension P of the parameter domain Θ is high, although techniques such as Smolyak sparse grids have recently been utilized to reduce the number of non-parametric problems [22, Sec. 7.2]. The disadvantage of the Galerkin method, on the other hand, is that almost the whole problem must be re-written and existing codes cannot be easily used in actual computations. Thus, in this context the Galerkin method is sometimes called *intrusive* approach.

For the Galerkin method, we need a finite-dimensional subspace of the function space $V \otimes L^2_w(\Theta)$. Let this subspace be spanned by $\{\phi_i\}_{i=1}^M \otimes \{\varphi_i\}_{j=1}^N$. The numerical solution $u_{M,N}(t)$ is now sought as a linear combination of those basis functions. Substituting

$$u_{M,N}(t) = \sum_{i=1}^{M} \sum_{j=1}^{N} \hat{u}_{i,j}(t) \phi_i(\boldsymbol{x}) \varphi_j(\boldsymbol{\vartheta}) \tag{2.7}$$

into the equation (2.6) and requiring the equality to hold for all basis functions results in a system of MN ordinary differential equations, namely

$$\partial_t \sum_{i=1}^{M} \sum_{j=1}^{N} R_{i,j,k,l} \hat{u}_{i,j}(t) + \sum_{i=1}^{M} \sum_{j=1}^{N} A_{i,j,k,l} \hat{u}_{i,j}(t) = F_{k,l}(t) + G_{k,l}(t), \tag{2.8}$$

where $k = 1, \ldots, M$ and $l = 1, \ldots, N$. Here,

$$R_{i,j,k,l} \coloneqq (\phi_i \varphi_j, \phi_k \varphi_l)_{L^2_w(\Omega \times \Theta)} \tag{2.9}$$

and

$$A_{i,j,k,l} \coloneqq (a \nabla \phi_i \varphi_j, \nabla \phi_k \varphi_l)_{L^2_w(\Omega \times \Theta)} \tag{2.10}$$

define the mass and stiffness matrices, respectively, and the vectors on the right hand side are defined according to

$$F_{k,l}(t) \coloneqq (f, \phi_k \varphi_l)_{L^2_w(\Omega \times \Theta)} \tag{2.11}$$

and

$$G_{k,l}(t) \coloneqq (g, \phi_k \varphi_l)_{L^2_w(\Gamma_{\mathrm{N}} \times \Theta)}. \tag{2.12}$$

The initial values for $\hat{u}_{i,j}$ are solved from the system

$$\sum_{i=1}^{M} \sum_{j=1}^{N} R_{i,j,k,l} \hat{u}_{i,j}(0) = (u_0, \phi_k \varphi_l)_{L^2_w(\Omega \times \Theta)}, \tag{2.13}$$

where again $k = 1, \ldots, M$ and $l = 1, \ldots, N$.

What remains is to choose the basis functions ϕ_i and φ_j and also the weight function w. In addition, a representation for the diffusivity a has to be chosen such that the requirement (2.3) is satisfied. The straightforward representation

$$a(\boldsymbol{x}, \boldsymbol{\vartheta}) = \sum_{p=1}^{P} \vartheta_p \psi_p(\boldsymbol{x}) \tag{2.14}$$

may or may not stay positive in the domain Ω. Previous works have largely concentrated on orthogonal basis functions ψ_p. Those have some nice properties and in stochastic settings they often correspond to so called Karhunen–Loève eigenfunctions [17, Sec. 2.1]. However, orthogonal functions do not

naturally satisfy the condition (2.3) and thus one must resort to a mapping s such as

$$s(y) = \exp(y), \quad s(y) = y^2 + a_{\min} \quad \text{or} \quad s(y) = y + C,$$

where y is the expansion (2.14) and C is a sufficiently large constant, in order to guarantee the positivity. Another option would be to restrict the domain Θ such that all vectors $\boldsymbol{\vartheta} \in \Theta$ yield a positive diffusivity, but this would result in difficulties compared to the simplest case of Θ being a hyperrectangle. In this thesis, we will choose functions ψ_p and a hyperrectangle Θ such that the requirement (2.3) is satisfied without an extra mapping s. Then the integral (2.10) becomes

$$A_{i,j,k,l} = \sum_{p=1}^{P} (\iota_p \psi_p \nabla \phi_i \varphi_j, \nabla \phi_k \varphi_l)_{L^2_w(\Omega \times \Theta)}, \tag{2.15}$$

where the projection $\iota_p \colon \Theta \to \mathbf{R}$ for $1 \leq p \leq P$ is defined by $\iota_p(\boldsymbol{\vartheta}) = \vartheta_p$. While some theoretical properties may be lost, the sparsity of the system that results from properly chosen non-orthogonal functions ψ_p significantly improves the computational potency, at least compared to the exponential mapping $s(y) = \exp(y)$. We will discuss the choice of both spatial function families $\{\phi_i\}_{i=1}^M$ and $\{\psi_p\}_{p=1}^P$ in the next section, whereas section 2.4 is devoted to discussion regarding the choices of the parametric basis functions $\{\varphi_j\}_{j=1}^N$ and the weight function w.

2.3 Spatial basis functions

We will use the finite element method (FEM) for the discretization of the space V that was defined in equation (2.5). For simplicity, let us first assume pure Neumann boundary conditions (i.e., $\Gamma_{\mathrm{D}} = \emptyset$) and piecewise linear basis functions $\{\phi_i\}_{i=1}^M$. The finite element mesh then consists of nodes $\{\xi_i\}_{i=1}^M \subset \overline{\Omega}$ and the basis functions satisfy $\phi_i(\xi_j) = \delta_{i,j}$. We denote the family of elements by $\{e_i\}_{i=1}^m \subset 2^\Omega$ so that in one spatial dimension we have $m = M - 1$.

Choosing the representation for the diffusivity is more interesting. As promised in the previous section, the functions ψ_p are chosen such that the strictly positive diffusivity admits a representation (2.14) where the ranges of the coefficients ϑ_p are independent and hence form a hyperrectangle. In practice, this means that each function ψ_p is non-negative on Ω and any linear combination with coefficient vector $\boldsymbol{\vartheta} \in \Theta$ yields a diffusivity which satisfies the condition (2.3). One obvious candidate is to use FEM basis functions and choose $\psi_i = \phi_i$ for $i = 1, \ldots, M = P$ with $\Theta = (a_{\min}, a_{\max})^M$. However, the

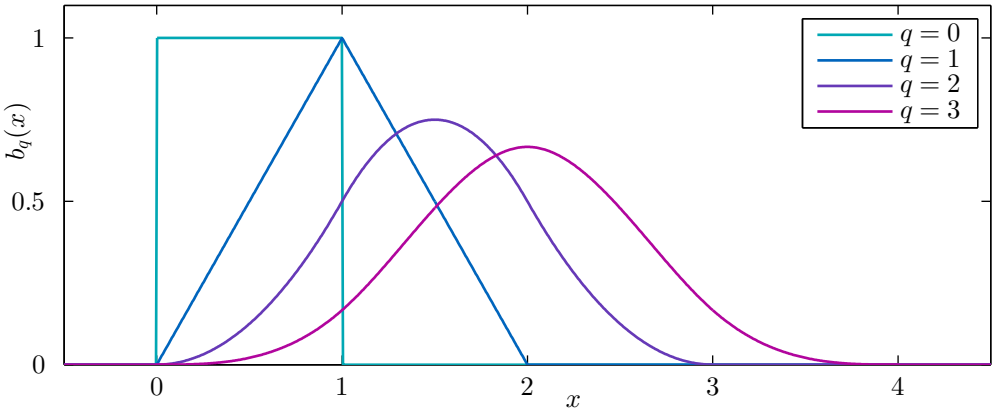

Figure 2.1: Standard uniform B-splines b_q for $0 \leq q \leq 3$.

number of solution basis functions ϕ_i typically needs to be larger than the number of parameter variables, that is, $M > P$. We will see that decreasing the number of parameters can substantially reduce the computational workload. Regarding inverse problems, the small number of parameter variables can also be used as a regularization technique, as discussed in section 4.2.

The condition $P = M$ is by no means necessary and thus we can construct the linear elements for the diffusivity independently on the solution mesh. Moreover, since the functions ψ_i are not differentiated in our problem and no boundary conditions are set, the benefits of a FEM basis are not so evident. Therefore, one may represent the diffusivity by piecewise constant indicator functions instead.

As a generalization of indicator functions and piecewise linear FEM basis functions, we can also employ B-splines. Recall that a univariate spline of degree q, where $q \in \mathbf{N}_0$, is a piecewise polynomial which is $q - 1$ times continuously differentiable. Here, $q = 0$ corresponds to a discontinuous function. As a special case of splines, the standard uniform B-spline of degree q is defined as a convolution

$$b_q(x) \coloneqq \int_{-\infty}^{\infty} b_{q-1}(x - y) b_0(y) \, \mathrm{d}y,$$

where $b_0(x)$ is the indicator function of the interval $[0, 1)$. Alternative but equivalent definitions exist [12, Chap. 3]. The first few standard uniform B-splines are illustrated in figure 2.1. It follows from the definition that the derivatives can show discontinuities only when x is an integer. Another important property is that any given polynomial of degree at most q can be represented as an infinite sum of shifted uniform B-splines $\{b_q(x - k)\}_{k \in \mathbf{Z}}$. Actually, any spline of degree q that is smooth between integers can be rep-

resented by the set $\{b_q(x-k)\}_{k\in\mathbf{Z}}$, which therefore is a basis for the corresponding function space [12, Sec. 3.4]. Moreover, the normalization is such that for any $q \in \mathbf{N}_0$ and $x \in \mathbf{R}$ we have

$$\sum_{k\in\mathbf{Z}} b_q(x-k) = 1,$$

that is, the shifted uniform B-splines form a partition of unity.

Instead of integers k, we can more generally consider a sequence of knots $\{\zeta_k\}_{k\in\mathbf{Z}} \subset \mathbf{R}$ which are uniformly spaced along the real axis and ordered such that the distance $\Delta\zeta \coloneqq \zeta_{k+1} - \zeta_k > 0$ is constant. The transformed uniform B-splines then become

$$\tilde{b}_{q,k}(x) \coloneqq b_q\left(\frac{x-\zeta_k}{\Delta\zeta}\right).$$

It is also possible to construct non-uniform B-splines for an arbitrary knot sequence such that most of the useful properties are preserved. Indeed, for $q \leq 1$ these are just the indicator functions and piecewise linear basis functions of an arbitrary (infinite) one-dimensional finite element mesh and similar multivariate B-splines can be easily defined in higher dimensions by using standard FEM techniques. However, univariate B-splines for $q > 1$ and especially the construction of high-degree multivariate B-splines as tensor products of univariate B-splines are most easily done by using uniform knot sequences.

A univariate B-spline of order q is positive between and only between the knots ζ_k and ζ_{k+q+1} for some $k \in \mathbf{Z}$. Consequently, the number of splines in a set $\{\tilde{b}_{q,k}\}_{k\in\mathbf{Z}}$ that do not vanish between two consecutive knots equals $q+1$, as depicted in figure 2.2 for $q = 2$. In order to obtain a basis for polynomials restricted on a bounded interval $\Omega \subset \mathbf{R}$ and having a degree q or less, it thus suffices to have $q+1$ splines and $2q+2$ knots from which 2 are at the boundaries of Ω and the rest lie on $\mathbf{R} \setminus \overline{\Omega}$. Increasing the number of interior knots $\zeta_k \in \Omega$ produces a spline space with dimension $P > q+1$ and results in better approximation properties. More splines also means narrower splines and eventually some splines will be compactly supported on Ω. Another way to improve the approximation power is to increase the degree of splines, which for fixed P essentially means that some interior knots are moved to the exterior. If the function to be approximated is smooth, increasing the degree instead of adding more splines may result in faster convergence or in some cases even an exact representation.

Constructing d-variate B-splines as tensor products of univariate splines is simple. The resulting splines inherit many properties such as the ability

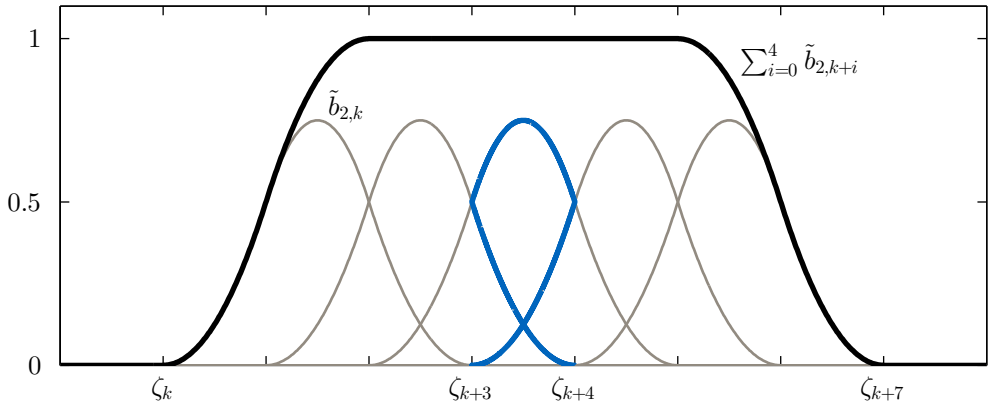

Figure 2.2: Quadratic B-splines $\tilde{b}_{2,k}, \ldots, \tilde{b}_{2,k+4}$ for some unspecified knot sequence $\{\zeta_k\}_{k \in \mathbf{Z}}$. The sum of splines, shown in black, is one between the knots ζ_{k+2} and ζ_{k+5}. Supported between the knots ζ_{k+3} and ζ_{k+4} are three splines, as visualized in blue.

to represent certain polynomials [12, Chap. 4]. In particular, tensor product splines can be scaled such that they form a partition of unity. In what follows, we will assume that this scaling has been done. Besides tensor product splines, any d-dimensional indicator function partitioning, as well as piecewise linear FEM bases, define valid splines and for those the unity scaling holds naturally. As the dimension d increases, the percentage of splines that are near the boundary and thus only partially supported grows. This holds especially for tensor product splines on a non-rectangular domain. To overcome possible problems, so called web-splines have been proposed [12, Sec. 4.4], but they are beyond the scope of this thesis.

Returning to choosing the representation for the diffusivity a, we see that one convenient choice for the functions $\{\psi_p\}_{p=1}^{P}$ is a set of B-splines. Because the splines sum up to one, the parameter domain, or the domain of the spline coefficients, can be taken as a Cartesian product $\Theta = (a_{\min}, a_{\max})^P$, where $a_{\min}$ and $a_{\max}$ define the minimum and maximum possible values of the diffusivity as in the condition (2.3). Those limits are rather artificial, because only the existence of the lower and upper bounds is required and the actual values may be irrelevant or unknown. Moreover, if the diffusivity function is known to be skew or otherwise irregular, we could consider a parameter domain of the form

$$\Theta = I_1 \times \cdots \times I_P, \tag{2.16}$$

where $I_p \subset \mathbf{R}$ are intervals for $1 \leq p \leq P$. A hyperrectangle of the form

(2.16) can be converted back to the form I^P by considering diffusivity

$$a(\boldsymbol{x}, \boldsymbol{\vartheta}) = c(\boldsymbol{x}) + \sum_{p=1}^{P} \vartheta_p \kappa_p \psi_p(\boldsymbol{x}), \tag{2.17}$$

where c is a "background diffusivity" and κ_p, $1 \leq p \leq P$, are positive scaling constants. By using affine transformations of the type (2.17), it is also possible to have parameter domains containing negative values. However, substituting (2.17) into the integral (2.10) results in a slightly different expression than the sum (2.15). We will briefly return to this issue later in this thesis.

Finally, let us discuss more about discretizing the space V. If homogeneous Dirichlet boundary conditions are present, the number of (nonzero) basis functions ϕ_i decreases and thus the number of nodes ξ_i becomes greater than M. The implementation of homogeneous Dirichlet boundary conditions is most straightforwardly done by first constructing the discrete operators for pure Neumann case and then removing the trivial parts from the resulting matrices (or tensors) and discarding the corresponding elements $\hat{u}_{i,j}$ from the unknown.

Using other than linear elements is at least conceptually straightforward. However, high-order elements provide benefits only when the diffusivity is smooth enough inside each element. This suggests that refining the discrete solution space should be done taking into account the degree of splines and the location of knots with respect to the element boundaries. For simplicity, we resort to linear elements. Note that due to the indefiniteness, the standard high-order FEM basis functions cannot easily be used to represent the diffusivity.

2.4 Basis functions for parameter domain

In this section, we discretize the space $L^2_w(\Theta)$ by introducing the basis fuctions $\{\varphi_j\}_{j=1}^N$. We assume that the hyperrectangular parameter domain can be written as $\Theta = I^P$ for some interval $I \subset \mathbf{R}$, but generalizing the methods to the case (2.16) is relatively straightforward. Because the dimension P is in general quite large, a finite element discretization would result in a very large number of elements. Therefore, instead of finite element method, we apply single-domain spectral method. To that end, we need basis functions that are orthogonal with respect to some positive weight function w, that is,

$$(\varphi_j, \varphi_l)_{L^2_w(\Theta)} = 0$$

whenever $j \neq l$. These basis functions are not differentiated in the equation (2.6) and thus the loss of sparsity, which is a typical issue in spectral methods, can be avoided.

Arguably the most natural way to construct multivariate orthogonal functions in a hyperrectangular domain is to form products of univariate orthogonal functions. Each univariate family is then orthogonal with respect to some weight function $\bar{w}_p \colon I \to \mathbf{R}_+$. For simplicity, we assume that all weight functions are the same, that is, $\bar{w}_1, \ldots, \bar{w}_P = \bar{w}$, and also that the orthogonal function families are the same, say $\{\bar{\varphi}_r\}_{r=0}^n$, for each direction. The obvious case we have in our mind is $\bar{\varphi}_r$ being orthogonal polynomials, whence the indexing starts from zero with $n \in \mathbf{N}_0$ denoting the maximum degree of a univariate polynomial. However, the present setting is more general and can include non-polynomial basis functions as well.

The multivariate weight function is a product

$$w(\boldsymbol{\vartheta}) = \prod_{p=1}^{P} \bar{w}(\vartheta_p).$$

For $1 \leq j \leq N$, we define the jth basis function as

$$\varphi_j(\boldsymbol{\vartheta}) = \prod_{p=1}^{P} \bar{\varphi}_{\Lambda_{j,p}}(\vartheta_p), \tag{2.18}$$

where $\boldsymbol{\Lambda} \in \mathbf{N}_0^{N \times P}$ is an *index matrix* which has N distinct rows and satisfies $0 \leq \Lambda_{j,p} \leq n$ for all its entries. It is easy to see that

$$(\varphi_j, \varphi_l)_{L^2_w(\Theta)} = \prod_{p=1}^{P} (\bar{\varphi}_{\Lambda_{j,p}}, \bar{\varphi}_{\Lambda_{l,p}})_{L^2_{\bar{w}}(I)} \tag{2.19}$$

vanishes whenever $j \neq l$, and if the univariate functions are orthonormal, then the multivariate functions are normalized as well.

Let us consider univariate orthogonal polynomials and their approximation properties in more detail. Mapping the weight function and the polynomials from a bounded interval to another is elementary and thus we can consider the standard interval $I = (-1, 1)$. Once the weight function $\bar{w}$ is chosen, the polynomials $\{\bar{\varphi}_r\}_{r=0}^n$ can always be constructed for example by the Gram–Schmidt process, assuming that the integrals

$$\int_I \vartheta^k \bar{w}(\vartheta) \, \mathrm{d}\vartheta$$

converge for sufficiently many $k \in \mathbf{N}_0$. For typical weight functions considered here, those integrals converge for all nonnegative integers k. In the

stochastic framework, the weight function is commonly chosen according to a density function of some probability distribution. Another and often not contradictory criterion is based on the convergence properties of the corresponding polynomial approximation. In the sense of the norm $\|\cdot\|_{L^2_{\bar{w}}(I)}$, the best nth degree approximation f_n of a function $f \in L^2_{\bar{w}}(I)$ is obtained by orthogonal projection onto the space spanned by $\{\bar{\varphi}_r\}_{r=0}^n$. It is known that the optimal asymptotic convergence rate for the error $\|f_n - f\|_{L^2_{\bar{w}}(I)}$ is achieved when using the Jacobi weight function

$$\bar{w}(\vartheta) = (1-\vartheta)^\alpha (1+\vartheta)^\beta, \qquad \alpha, \beta > -1. \tag{2.20}$$

This is a consequence of a certain singular Sturm–Liouville eigenvalue problem [5, Sec. 5.2]. Since the degree n is usually relatively low in parametric and stochastic equations, the asymptotic properties are not always of main interest, but in some cases the approximation may fail completely if the weight function is poorly chosen [8].

The polynomials that are orthogonal with respect to the weight function (2.20) are called Jacobi polynomials. Of course, a notable special case follows from $\alpha = \beta = 0$, which corresponds to unweighted L^2 inner products and the Legendre polynomials $\bar{\varphi}_r = L_r$. A computationally useful recurrence relation for the Legendre polynomials is

$$L_{r+1}(\vartheta) = \frac{2r+1}{r+1}\vartheta L_r(\vartheta) - \frac{r}{r+1}L_{r-1}(\vartheta),$$

where $L_0(\vartheta) = 1$ and $L_1(\vartheta) = \vartheta$ [5, Sec. 2.3][10]. Regarding the multivariate polynomials as products of univariate orthogonal polynomials as in equation (2.19), it is usually convenient to normalize the polynomials such that

$$(\bar{\varphi}_r, \bar{\varphi}_s)_{L^2_{\bar{w}}(I)} = \delta_{r,s}.$$

In addition, normalizing the weight function according to $\|\bar{w}\|_{L^1(I)} = 1$ may provide computational benefits. This of course requires the re-normalization of the polynomials.

If the interval I is large, it may be tempting to consider an unbounded domain $I = (0, \infty)$ instead. The corresponding Sturm–Liouville problem yields the Laguerre weight function $\bar{w}(\vartheta) = \mathrm{e}^{-\vartheta}$ and the Laguerre polynomials [5, Sec. 2.6]. Although the condition (2.3) is violated, it is likely that the numerical results stay close to ones that correspond to a large but bounded interval [2, Chap. 17].

Now that we have established means to construct univariate orthogonal polynomials, we consider how to choose the multivariate basis $\{\varphi_j\}_{j=1}^N$ and

the dimension N thereof. As in equation (2.18), the construction of multivariate functions reduces to choosing the $N \times P$ index matrix $\boldsymbol{\Lambda}$. The matrix does not depend on the underlying univariate functions and thus the following constructions can be equally applied for polynomials and other orthogonal functions. The largest possible index matrix corresponds to all combinations of $n+1$ univariate functions so that the number of rows is $N = (n+1)^P$. The resulting tensor product space is

$$\{\varphi_j\}_{j=1}^{N} = \left\{ \prod_{p=1}^{P} \bar{\varphi}_{r_p}(\vartheta_p) \;\middle|\; 0 \leq r_p \leq n,\ p = 1, \ldots, P \right\}. \tag{2.21}$$

Even for moderate values of n and P, the number of functions in the space (2.21) grows very fast. A common alternative is the total degree (TD) space, defined as

$$\{\varphi_j\}_{j=1}^{N} = \left\{ \prod_{p=1}^{P} \bar{\varphi}_{r_p}(\vartheta_p) \;\middle|\; \sum_{p=1}^{P} r_p \leq n \right\}. \tag{2.22}$$

This corresponds to an index matrix having an upper bound n for its row sums. It can be shown that the number of such rows is

$$N = N_{\mathrm{TD}}(P, n) \coloneqq \binom{P+n}{P} = \frac{(P+n)!}{P!\, n!}. \tag{2.23}$$

In terms of polynomials, the TD space contains only those multivariate polynomials whose total degree is less than or equal to n, resembling the terms in a truncated Taylor series.

If the coefficients ϑ_p are considered to be of unequal importance, the function space can be adjusted accordingly. In [6], for example, an anisotropic space based on the norms $\|\psi_p\|_{L^\infty(\Omega)}$ of the diffusivity functions is proposed. In our setting, this could be relevant when the degree of splines is $q > 1$ as some of the splines would then attain their maximum value outside the domain Ω. Several different polynomial spaces are presented in [4]. For simplicity, however, we will rely on the total degree space defined in equation (2.22).

Chapter 3

Computations

In this chapter, we show how the matrix equation corresponding to the discretized parametric diffusion equation can be constructed and solved. We focus on issues that arise due to the large number of unknowns in the linear system. In particular, the sparsity of the system is analyzed in detail. Although code snippets are omitted, a careful reader should be able to reproduce the experiments of this thesis by using an appropriate high-level programming language such as Matlab.

3.1 Constructing the system

We next discuss the construction of the matrices and vectors appearing in the linear system (2.8). As can be seen, the mass and stiffness matrices are actually defined as 4-way tensors [16], while the unknown and the right hand side consist of matrices, but unfolding these objects by matricization and vectorization is straightforward. In principle, the order in which to arrange the elements does not matter, as long as the ordering is consistent, because the resulting matrices and vectors will be permuted anyway when solving the system. However, it is convenient to define the vector of unknown coefficients as

$$\hat{\boldsymbol{u}}(t) \coloneqq [\hat{u}_{1,1}(t), \hat{u}_{2,1}(t), \ldots, \hat{u}_{M,1}(t), \hat{u}_{1,2}(t), \ldots, \hat{u}_{M-1,N}(t), \hat{u}_{M,N}(t)]^{\mathrm{T}} \quad (3.1)$$

and assemble the other arrays accordingly.

It is easy to see that the integrals in equations (2.11)–(2.13) can be decomposed such that

$$(\cdot, \phi_k \varphi_l)_{L^2_w(\Omega\times\Theta)} = (\cdot, \phi_k)_{L^2(\Omega)} (1, \varphi_l)_{L^2_w(\Theta)}. \quad (3.2)$$

A common case is that φ_1 is a constant L^2_w-normalized function, or more precisely

$$\varphi_1(\boldsymbol{\vartheta}) = \|\bar{w}\|_{L^1(I)}^{-P/2}, \qquad \boldsymbol{\vartheta} \in \Theta,$$

where the univariate weight function $\bar{w}$, the interval I and the domain $\Theta = I^P$ are as in section 2.4. Due to orthogonality, the factor $(1, \varphi_l)_{L^2_w(\Theta)}$ in equation (3.2) then vanishes for all but $l = 1$, for which it results in

$$\mu \coloneqq (1, \varphi_1)_{L^2_w(\Theta)} = \|\bar{w}\|_{L^1(I)}^{P/2}.$$

Hence, by using the ordering which is consistent with the definition (3.1), the right hand side vectors (2.11) and (2.12) become

$$\hat{\boldsymbol{f}}(t) \coloneqq \mu[(f(t), \phi_1)_{L^2(\Omega)}, \ldots, (f(t), \phi_M)_{L^2(\Omega)}, 0, \ldots, 0]^{\mathrm{T}} \tag{3.3}$$

and

$$\hat{\boldsymbol{g}}(t) \coloneqq \mu[(g(t), \phi_1)_{L^2(\Gamma_{\mathrm{N}})}, \ldots, (g(t), \phi_M)_{L^2(\Gamma_{\mathrm{N}})}, 0, \ldots, 0]^{\mathrm{T}},$$

respectively, and the initial condition vector on the right hand side of equation (2.13) is

$$\hat{\boldsymbol{u}}_0 \coloneqq \mu[(u_0, \phi_1)_{L^2(\Omega)}, \ldots, (u_0, \phi_M)_{L^2(\Omega)}, 0, \ldots, 0]^{\mathrm{T}}. \tag{3.4}$$

Those are just appropriately scaled vectors that arise from standard non-parametric FEM discretization, extended by $M(N-1)$ zeros. If the basis functions do not contain a constant function or if the initial condition, forcing term or boundary terms are parametrized similar to the diffusivity, the vectors (3.3)–(3.4) in general contain less zeros. Even in that case, the integrals would probably be easy to compute.

From now on, we assume that the basis $\{\varphi_j\}_{j=1}^N$ is orthonormal, that is, $(\varphi_j, \varphi_l)_{L^2_w(\Theta)} = \delta_{j,l}$. Based on that, the integral in (2.9) can be written as

$$(\phi_i \varphi_j, \phi_k \varphi_l)_{L^2_w(\Omega\times\Theta)} = (\phi_i, \phi_k)_{L^2(\Omega)} (\varphi_j, \varphi_l)_{L^2_w(\Theta)} = (\phi_i, \phi_k)_{L^2(\Omega)} \delta_{j,l}. \tag{3.5}$$

The mass matrix $\boldsymbol{R} \in \mathbf{R}^{MN\times MN}$ can then be defined as a block diagonal matrix

$$\boldsymbol{R} \coloneqq \boldsymbol{I}_{N\times N} \otimes \boldsymbol{R}^{\bullet} \tag{3.6}$$

where $\otimes$ denotes the Kronecker product [21] and $\boldsymbol{R}^{\bullet} \in \mathbf{R}^{M\times M}$, defined by the entries $R^{\bullet}_{i,k} = (\phi_i, \phi_k)_{L^2(\Omega)}$, is the symmetric mass matrix of the corresponding non-parametric problem. It is easy to verify that the elements in $\boldsymbol{R}$ are now ordered in a way that $\boldsymbol{R}\hat{\boldsymbol{u}}(t)$ is equivalent to the first double sum in equation (2.8).

Let us then discuss the construction of the stiffness matrix $\boldsymbol{A} \in \mathbf{R}^{MN \times MN}$, which appears in both parabolic and elliptic parametric equations. In general, computing the integrals in (2.10) may be a bit complicated (see e.g. [18]), but in our case it suffices to handle integrals of the form

$$(\iota_p \psi_p \nabla \phi_i \varphi_j, \nabla \phi_k \varphi_l)_{L^2_w(\Omega \times \Theta)} = (\psi_p \nabla \phi_i, \nabla \phi_k)_{L^2(\Omega)} (\iota_p \varphi_j, \varphi_l)_{L^2_w(\Theta)}, \tag{3.7}$$

which arise from equation (2.15). Similar to the mass matrix, the integrals (3.7) for $i, k = 1, \ldots, M$ and $j, l = 1, \ldots, N$ can be stored in a matrix by applying a tensor product

$$\boldsymbol{A}^{(p)} \coloneqq \boldsymbol{Y}^{(p)} \otimes \boldsymbol{X}^{(p)}, \tag{3.8}$$

where

$$X^{(p)}_{i,k} \coloneqq (\psi_p \nabla \phi_i, \nabla \phi_k)_{L^2(\Omega)} \tag{3.9}$$

and

$$Y^{(p)}_{j,l} \coloneqq (\iota_p \varphi_j, \varphi_l)_{L^2_w(\Theta)} \tag{3.10}$$

define symmetric matrices of sizes $M \times M$ and $N \times N$, respectively. Following equation (2.15), the stiffness matrix is then defined as

$$\boldsymbol{A} \coloneqq \sum_{p=1}^{P} \boldsymbol{A}^{(p)}. \tag{3.11}$$

The matrix $\boldsymbol{X}^{(p)} \in \mathbf{R}^{M \times M}$ is nothing but the FEM stiffness matrix of the corresponding non-parametric equation with diffusivity ψ_p. Since a spline ψ_p is in general not supported on the whole domain Ω, the property (2.3) is not satisfied, but the computation of the entries of $\boldsymbol{X}^{(p)}$ can be carried out by using standard methods.

Computing the elements of the matrix $\boldsymbol{Y}^{(p)} \in \mathbf{R}^{N \times N}$ requires a bit more attention. First, we note that

$$Y^{(p)}_{j,l} = (\iota_p \varphi_j, \varphi_l)_{L^2_w(\Theta)} = C_{\Lambda_{j,p}, \Lambda_{l,p}} B^{(p)}_{j,l}, \tag{3.12}$$

where

$$B^{(p)}_{j,l} \coloneqq \prod_{\substack{1 \le q \le P \\ q \neq p}} (\bar{\varphi}_{\Lambda_{j,q}}, \bar{\varphi}_{\Lambda_{l,q}})_{L^2_{\bar{w}}(I)} \tag{3.13}$$

for $1 \le j, l \le N$ defines a large matrix $\boldsymbol{B}^{(p)} \in \mathbf{R}^{N \times N}$ and the smaller matrix $\boldsymbol{C} \in \mathbf{R}^{(n+1) \times (n+1)}$ contains the univariate integrals

$$C_{r,s} \coloneqq \int_I \vartheta \bar{\varphi}_r(\vartheta) \bar{\varphi}_s(\vartheta) \bar{w}(\vartheta) \, \mathrm{d}\vartheta \tag{3.14}$$

for $r, s = 0, \ldots, n$. The product (3.13) is 1 if $j = l$ or if the jth and lth rows of the index matrix $\boldsymbol{\Lambda}$ differ only at the pth element. Otherwise, the product vanishes and $B^{(p)}_{j,l} = 0$. Thus, the remaining task is to assign the values in $\boldsymbol{C}$ into the much larger but very sparse matrix $\boldsymbol{Y}^{(p)}$.

Clearly, both $\boldsymbol{B}^{(p)}$ and $\boldsymbol{C}$ are symmetric matrices, but for polynomial functions $\bar{\varphi}_r$ we have the additional property that $\boldsymbol{C}$ is triagonal. This further increases the sparsity of the matrix $\boldsymbol{Y}^{(p)}$. The tridiagonality follows from the fact that an orthogonal polynomial is orthogonal to *all* polynomials of lower degree and thus the integral in (3.14) vanishes whenever $|r - s| > 1$. Moreover, if the weight function $\bar{w}$ is symmetric about the origin, the diagonal of $\boldsymbol{C}$ is zero. By the change of variables, we see that any shifted and scaled Jacobi weight function with $\alpha = \beta$ in (2.20) results in $\boldsymbol{C}$ with a constant diagonal. We deduce that when using multivariate Legendre polynomials, the number of distinct nonzero values in the matrix $\boldsymbol{C}$ and also in the matrix $\boldsymbol{Y}^{(p)}$ is at most $n + 1$.

Typically, the number of univariate functions is small and the matrix $\boldsymbol{C}$ is easy to construct. In the polynomial case, for example, the integrals (3.14) can be computed numerically but exactly by using a Gaussian quadrature rule with $n + 1$ nodes [10, Sec. 1.4]. On the other hand, the diagonal of the matrix $\boldsymbol{B}^{(p)}$ contains only ones. The off-diagonal elements of $\boldsymbol{B}^{(p)}$ can be constructed by first extracting the non-unique rows from the matrix which is obtained by removing the pth column from the index matrix $\boldsymbol{\Lambda}$. By keeping track of the indices of the rows and their duplicates, all possible index pairs (j, l), for which $B^{(p)}_{j,l} = 1$, can be found. Note that the construction of the matrices $\boldsymbol{B}^{(p)}$, $\boldsymbol{C}$ and $\boldsymbol{Y}^{(p)}$ is quite general and does not resort to any particular kind of function space. However, for different values of p, the matrices $\boldsymbol{Y}^{(p)}$ share many common properties only when an isotropic function space such as the tensor product space (2.21) or the total degree space (2.22) is used.

3.2 Sparsity

The dimension MN of the discrete system (2.8) can be very large. As an example, a crude discretization in two spatial dimensions with $M = 25^2$ FEM basis functions ϕ_i and $P = 4^2$ diffusivity functions ψ_p would result in $MN \approx 6 \cdot 10^5$ if a total degree space with $n = 3$ in (2.23) was used. Fortunately, the mass and stiffness matrices are very sparse, that is, they contain a lot of zeros. At least up to a certain limit, it is thus possible to store the matrices in the memory and carry out the computations efficiently, despite of the huge size. In this section, we will investigate and quantify the

sparsity of the system, whereas the actual solving issues, including the fill-in of the system, will be considered in the next section.

We denote the number of nonzero elements of any matrix $\boldsymbol{V}$ by $\operatorname{nnz}(\boldsymbol{V})$. It is obvious that the Kronecker product satisfies

$$\operatorname{nnz}(\boldsymbol{V} \otimes \boldsymbol{W}) = \operatorname{nnz}(\boldsymbol{V}) \operatorname{nnz}(\boldsymbol{W}).$$

In particular, for the mass matrix $\boldsymbol{R}$ defined in (3.6) we have

$$\operatorname{nnz}(\boldsymbol{R}) = N \operatorname{nnz}(\boldsymbol{R}^\bullet).$$

Another elementary fact is that

$$\operatorname{nnz}\left(\sum_i \boldsymbol{V}_i\right) \leq \sum_i \operatorname{nnz}(\boldsymbol{V}_i) \tag{3.15}$$

holds for any matrices $\boldsymbol{V}_i$ of the same size. Once we know the number of nonzero elements for matrices $\boldsymbol{X}^{(p)}$ and $\boldsymbol{Y}^{(p)}$, the inequality (3.15) provides an upper bound for the stiffness matrix $\boldsymbol{A}$ defined as the sum (3.11).

Analogous to the mass matrix $\boldsymbol{R}^\bullet$, we define the standard non-parametric stiffness matrix $\boldsymbol{A}^\bullet \in \mathbf{R}^{M \times M}$ according to

$$A^\bullet_{i,k} \coloneqq (\nabla \phi_i, \nabla \phi_k)_{L^2(\Omega)}. \tag{3.16}$$

In general, the matrices $\boldsymbol{R}^\bullet$ and $\boldsymbol{A}^\bullet$ have the same sparsity structure. For the spatial matrices defined in (3.9), we have $\operatorname{nnz}(\boldsymbol{X}^{(p)}) \leq \operatorname{nnz}(\boldsymbol{A}^\bullet)$ and the equality holds if the spline ψ_p is supported on the whole domain Ω. Note that because the splines ψ_p form a partition of unity, we always have

$$\sum_{p=1}^{P} \boldsymbol{X}^{(p)} = \boldsymbol{A}^\bullet. \tag{3.17}$$

We will soon make use of an auxiliary measure η defined as

$$\eta \coloneqq \frac{\sum_{p=1}^{P} \operatorname{nnz}(\boldsymbol{X}^{(p)})}{\operatorname{nnz}(\boldsymbol{A}^\bullet)}, \tag{3.18}$$

which clearly satisfies $1 \leq \eta \leq P$. The value of η measures how much the matrices $\boldsymbol{X}^{(p)}$ "overlap" in the sense of their sparsity structure. Largely, this reduces to measuring the overlapping of the splines. Increasing the degree q of splines makes them overlap more, but also the case $q = 0$, corresponding to piecewise constant and non-overlapping splines, results in $\eta > 1$ unless only one spline is employed. In order to decrease η, the supports of individual

splines should match with the FEM elements as much as possible. By using the notation introduced in section 2.3, we would like to have

$$\operatorname{supp}(\psi_p) \bigcap \overline{\Omega} = \bigcup_{i \in \mathcal{I}_p} \overline{e_i} \tag{3.19}$$

for some index set $\mathcal{I}_p \subseteq \{1, \ldots, m\}$ and for $1 \leq p \leq P$. In one spatial dimension, for example, this can be achieved by choosing the interior and boundary knots to be a subset of the FEM nodes, that is, $\{\zeta_k\} \cap \overline{\Omega} \subseteq \{\xi_k\}$. Although the requirement (3.19) is not strictly necessary, it does not only result in improved sparsity but also makes computing the integrals (3.9) easier.

Regarding the parametric matrices $\boldsymbol{Y}^{(p)}$, we mainly restrict our discussion to matrices that result from a total degree space (2.22). In that case, all matrices $\boldsymbol{Y}^{(p)}$ for $p = 1, \ldots, P$ have the same number of nonzero elements.

Theorem. *Assume that the P-variate w-orthogonal functions $\{\varphi_j\}_{j=1}^{N}$ are constructed according to a total degree space* (2.22) *such that $N = N_{TD}(P, n)$ for some $n \in \mathbf{N}_0$ as in equation* (2.23)*. Assume further that the matrix $\boldsymbol{C}$, defined in* (3.14)*, is full. Then the number of nonzero elements in the matrix $\boldsymbol{Y}^{(p)}$, defined in* (3.10)*, is*

$$\operatorname{nnz}(\boldsymbol{Y}^{(p)}) = N + 2 \sum_{k=0}^{n-1} N_{TD}(P, k) = \left(1 + \frac{2n}{P+1}\right) N \tag{3.20}$$

for all $1 \leq p \leq P$.

Proof. Based on equation (3.12), the task reduces to determining the number of nonzero elements in the matrix $\boldsymbol{B}^{(p)}$. Furthermore, we already know that $\boldsymbol{B}^{(p)}$ is symmetric and its diagonal is full. Hence, it suffices to show that the product (3.13) vanishes for all but $\sum_{k=0}^{n-1} N_{\mathrm{TD}}(P, k)$ index pairs (j, l), where $j < l$. Without loss of generality, we can assume that the index matrix $\boldsymbol{\Lambda} \in \mathbf{R}^{N \times P}$ is arranged so that

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}^{(0)} \\ \vdots \\ \boldsymbol{\Lambda}^{(n)} \end{bmatrix},$$

where, for every $0 \leq r \leq n$, each row sum of the submatrix $\boldsymbol{\Lambda}^{(r)} \in \mathbf{N}_0^{N_r \times P}$ equals r. Clearly, $N_0 = N_{\mathrm{TD}}(P, 0) = 1$ and

$$N_r = N_{\mathrm{TD}}(P, r) - N_{\mathrm{TD}}(P, r-1)$$

for $r \geq 1$. A given submatrix $\boldsymbol{\Lambda}^{(r)}$ cannot have two rows that differ only at the pth element, because due to the row sum constraint the pth element is uniquely determined by the other values on the row. On the other hand, if the jth row of the matrix $\boldsymbol{\Lambda}$ is contained in $\boldsymbol{\Lambda}^{(r)}$, then every submatrix $\boldsymbol{\Lambda}^{(k)}$, where $r < k \leq n$, contains a row, say the lth row of $\boldsymbol{\Lambda}$, that equals the jth row of $\boldsymbol{\Lambda}$ except that the pth element is replaced by $\Lambda_{l,p} = \Lambda_{j,p} + k - r$. Each such row pair (j, l) contributes to the matrix $\boldsymbol{B}^{(p)}$ as in equation (3.13). Now the number of such pairs (j, l), for which $j < l$, equals

$$nN_0 + (n-1)N_1 + \ldots + 2N_{n-2} + N_{n-1} = \sum_{k=0}^{n-1} N_{\mathrm{TD}}(P, k) \tag{3.21}$$

as claimed. The second equality in (3.20), namely

$$\sum_{k=0}^{n-1} \binom{P+k}{P} = \frac{n}{P+1} \binom{P+n}{P},$$

is left for the reader. □

If the matrix $\boldsymbol{C}$ is tridiagonal, as it is in the polynomial case, then the integral (3.12), where j and l correspond to submatrices $\boldsymbol{\Lambda}^{(r)}$ and $\boldsymbol{\Lambda}^{(k)}$, respectively, vanishes whenever $k > r+1$. In that case, the sum (3.21) reduces to a telescoping sum

$$N_0 + N_1 + \ldots + N_{n-2} + N_{n-1} = N_{\mathrm{TD}}(P, n-1).$$

Thus, the number of nonzero elements in $\boldsymbol{Y}^{(p)}$ when using polynomials is

$$\mathrm{nnz}(\boldsymbol{Y}^{(p)}) = N + 2N_{\mathrm{TD}}(P, n-1) = \left(1 + \frac{2n}{P+n}\right) N. \tag{3.22}$$

Similar reasoning shows that if $\boldsymbol{C}$ is a banded matrix with s subdiagonals, then $\mathrm{nnz}(\boldsymbol{Y}^{(p)})$ is obtained by picking out the last s terms from the sum in equation (3.20).

Even though the number of nonzeros is the same for all $\boldsymbol{Y}^{(p)}$ when using a total degree space, the non-vanishing values are of course placed in different positions in the matrices. Indeed, apart from the diagonal, the index pairs (j, l) for which the integral (3.10) does not vanish are disjoint for $1 \leq p \leq P$. This is due to the fact that if the pth matrix $\boldsymbol{Y}^{(p)}$ has a nonzero value at position (j, l), then the jth and lth rows of the index matrix $\boldsymbol{\Lambda}$ can differ at most at the pth element. But if also $\boldsymbol{Y}^{(q)}$ has a nonzero value at (j, l), then those rows can differ at most at the qth element. This implies that $p = q$ or $j = l$, whence the positions of the non-diagonal elements in $\boldsymbol{Y}^{(p)}$ and $\boldsymbol{Y}^{(q)}$

do not coincide for $p \neq q$. Note that the previous fact holds for any set of orthogonal functions, not just for the total degree space.

We are now ready to establish a result for the number of nonzero elements in the stiffness matrix $\boldsymbol{A}$ defined in (3.11). Let us first assume that the matrix $\boldsymbol{C}$ has a full diagonal and hence by equation (3.12) the diagonal of $\boldsymbol{Y}^{(p)}$ is full as well. Now the sparsity structure of the diagonal blocks of $\boldsymbol{A}$ equals that of $\boldsymbol{A}^\bullet$, following from equation (3.17). The total number of nonzero elements in the diagonal blocks is thus simply $N\,\mathrm{nnz}(\boldsymbol{A}^\bullet)$. On the other hand, the disjointness of the off-diagonal elements in the matrices $\boldsymbol{Y}^{(p)}$ yields

$$\begin{aligned}\mathrm{nnz}(\boldsymbol{A}) &= N\,\mathrm{nnz}(\boldsymbol{A}^\bullet) + \sum_{p=1}^{P}(\mathrm{nnz}(\boldsymbol{Y}^{(p)}) - N)\,\mathrm{nnz}(\boldsymbol{X}^{(p)}) \\ &= \mathrm{nnz}(\boldsymbol{A}^\bullet)\Big(\eta\,\mathrm{nnz}(\boldsymbol{Y}^{(p)}) - N(\eta-1)\Big),\end{aligned} \tag{3.23}$$

where the last expression containing the quantity (3.18) is of course meaningless if $\mathrm{nnz}(\boldsymbol{Y}^{(p)})$ depends on p. Substituting the value (3.20), corresponding to a total degree space with general non-polynomial basis functions φ_j and a full $\boldsymbol{C}$, results in

$$\mathrm{nnz}(\boldsymbol{A}) = N\,\mathrm{nnz}(\boldsymbol{A}^\bullet)\Big(1 + \frac{2n\eta}{P+1}\Big). \tag{3.24}$$

For a total degree polynomial space and a tridiagonal $\boldsymbol{C}$, we obtain

$$\mathrm{nnz}(\boldsymbol{A}) = N\,\mathrm{nnz}(\boldsymbol{A}^\bullet)\Big(1 + \frac{2n\eta}{P+n}\Big), \tag{3.25}$$

which follows from (3.23) and (3.22). Obviously, (3.25) cannot be greater than (3.24).

Now suppose that $\boldsymbol{C}$ has an empty diagonal. As mentioned in the previous section, the diagonal vanishes if the weight function $\bar{w}$ is symmetric about the origin. We see from equation (3.12) that the diagonal of the matrix $\boldsymbol{Y}^{(p)}$ then becomes zero as well. Furthermore, other entries of $\boldsymbol{Y}^{(p)}$ are not affected, because if $B_{j,l}^{(p)}$ is nonzero and $j \neq l$, then the jth and lth rows of $\boldsymbol{\Lambda}$ differ exactly at the pth element, that is, $\Lambda_{j,p} \neq \Lambda_{l,p}$. In other words, the number of nonzero elements in the block matrix $\boldsymbol{A}$ decreases by $N\,\mathrm{nnz}(\boldsymbol{A}^\bullet)$ compared to the case where the diagonal of $\boldsymbol{C}$ is full. However, having a non-positive parameter domain and a symmetric weight function is possible only if the diffusivity is expressed as in equation (2.17), where c is positive on the whole domain Ω. Now the sum (2.15) must be replaced by

$$A_{i,j,k,l} = (c\nabla\phi_i\varphi_j, \nabla\phi_k\varphi_l)_{L^2_w(\Omega\times\Theta)} + \sum_{p=1}^{P}(\iota_p\kappa\psi_p\nabla\phi_i\varphi_j, \nabla\phi_k\varphi_l)_{L^2_w(\Omega\times\Theta)}, \tag{3.26}$$

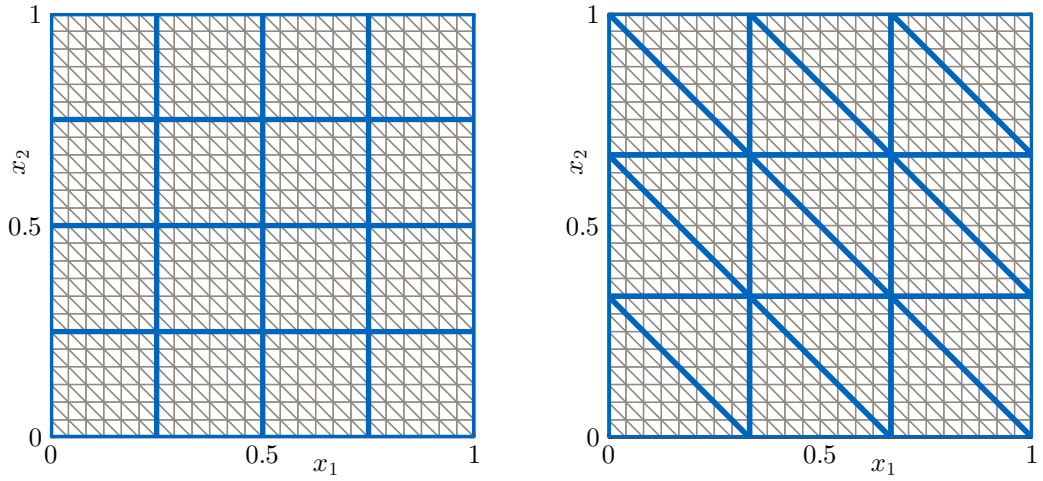

Figure 3.1: Finite element mesh with 625 nodes and 1152 triangles, superimposed with the supports of the diffusivity splines $\{\psi_p\}_{p=1}^{16}$. Left: piecewise constant splines with rectangular supports. Right: piecewise linear splines with FEM triangulation. Both satisfy the relation (3.19).

where the first term can be decomposed as

$$(c\nabla\phi_i\varphi_j, \nabla\phi_k\varphi_l)_{L^2_w(\Omega\times\Theta)} = (c\nabla\phi_i, \nabla\phi_k)_{L^2(\Omega)}\delta_{j,l}$$

similar to equation (3.5). Thus, the first term in (3.26) contributes to the sparsity structure of $\boldsymbol{A}$ exactly as $\boldsymbol{I}_{N\times N} \otimes \boldsymbol{A}^\bullet$. The resulting stiffness matrix then has the same properties as what was already obtained in the case where the affine mapping (2.17) was not present.

To get some idea of the numerical values of η and of the sparsity of $\boldsymbol{A}$, let us return to the example mentioned in the beginning of this section. To that end, we construct a finite element mesh on $\Omega = (0,1)^2$ with $m = 1152$ elements as illustrated in figure 3.1. Assuming pure Neumann boundary conditions, we then have $M = 25^5$. We use shifted Legendre polynomials $\{\varphi_j\}_{j=1}^N$, where $N = N_{\mathrm{TD}}(P,n) = 969$ for $P = 4^2, n = 3$, and experiment with two different sets of diffusivity functions $\{\psi_p\}_{p=1}^P$. The first set consists of piecewise constant splines such that each spline is supported precisely on some subset of the triangle family $\{e_i\}_{i=1}^m$, satisfying condition (3.19). We obtain

$$\eta = \frac{3472}{3025} \approx 1.15$$

as in equation (3.18). By using equation (3.25), the average number of nonzero entries per row in the stiffness matrix $\boldsymbol{A} \in \mathbf{R}^{MN\times MN}$ can then

be expressed as

$$\overline{\mathrm{nnz}}(\boldsymbol{A}) \coloneqq \frac{\mathrm{nnz}(\boldsymbol{A})}{MN} = \frac{3993657}{605625} \approx 6.59. \tag{3.27}$$

For the piecewise linear diffusivity functions shown on the right in figure 3.1, we obtain

$$\eta = \frac{9448}{3025} \approx 3.12$$

and

$$\overline{\mathrm{nnz}}(\boldsymbol{A}) = \frac{5822313}{605625} \approx 9.61.$$

If the diffusivity functions were supported on the whole domain Ω, yielding $\eta = P = 16$, the number of nonzero elements in $\boldsymbol{A}$ would become about threefold compared to the piecewise linear splines. Indeed, equation (3.25) shows that in this case we have

$$\overline{\mathrm{nnz}}(\boldsymbol{A}) = \frac{17741625}{605625} \approx 29.3.$$

3.3 Solving the system

The system of differential equations (2.8) can be compactly written as

$$\partial_t \boldsymbol{R}\hat{\boldsymbol{u}}(t) + \boldsymbol{A}\hat{\boldsymbol{u}}(t) = \hat{\boldsymbol{r}}(t), \tag{3.28}$$

where the right hand side is

$$\hat{\boldsymbol{r}}(t) \coloneqq \hat{\boldsymbol{f}}(t) + \hat{\boldsymbol{g}}(t)$$

and all other matrices and vectors are as in section 3.1. The steady-state problem concerns the equation

$$\boldsymbol{A}\hat{\boldsymbol{u}} = \lim_{t\to\infty} \hat{\boldsymbol{r}}(t), \tag{3.29}$$

which can be solved by using standard methods whenever the limit exists and $\Gamma_{\mathrm{D}} \neq \emptyset$. For pure Neumann boundary conditions, the stiffness matrix $\boldsymbol{A}$ is singular and some additional work is required.

Iterative methods for equation (3.29) essentially consist of sparse matrix-vector multiplications and finding a good preconditioner may be crucial for fast convergence [11, Chap. 11]. Conversely, direct methods involve factorizing the stiffness matrix into two triangular matrices, and reordering the rows and columns prior to the factorization is important in order to reduce the fill-in, or the loss of sparsity, in the system. In this section, we concentrate

on the time-dependent problem (3.28), but similar issues arise for both the parabolic system (3.28) and the elliptic equation (3.29).

When using an explicit time integration method for equation (3.28), the integration essentially reduces to computations which can be formally written as

$$\boldsymbol{v}^{(1)} = \boldsymbol{R}^{-1}(-\boldsymbol{A}\boldsymbol{v}^{(2)} + \boldsymbol{v}^{(3)}), \tag{3.30}$$

where the vectors $\boldsymbol{v}^{(i)}$ vary between the iterations and also between the stages within one iteration. In practice, the symmetric and positive-definite mass matrix $\boldsymbol{R}$ is decomposed as $\boldsymbol{R} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$, where $\boldsymbol{L} =: \mathrm{chol}(\boldsymbol{R})$ is the Cholesky factor of $\boldsymbol{R}$ [11, Sec. 4.2]. The resulting triangular systems are then easy to solve. The factorization and solving can be done block-wise based on the definition (3.6) of the mass matrix. In one spatial dimension, the matrix $\boldsymbol{R}$ is tridiagonal and thus the Cholesky factor does not suffer from fill-in, but in higher dimensions, depending on the structure and indexing of the FEM mesh, it may be relevant to permute the mass matrix, or equivalently one of its diagonal blocks.

It is known that second order partial differential equations impose severe restrictions for the time step size of an explicit time integration method. Based on our numerical experiments, the maximum (absolute) eigenvalue of the stiffness matrix $\boldsymbol{A}$ seems to be slightly smaller but still of the same order of magnitude as that of the non-parametric matrix $\boldsymbol{A}^{\bullet}$ (see Eq. (3.16)), if the latter is scaled by the maximum possible value of the diffusivity. Thus, the stability of the explicit iterative scheme (3.30) may become an issue and an implicit alternative is often wanted. Here, we consider the implicit midpoint rule, or the Crank–Nicolson method, which for a constant time step $\tau > 0$ is written as

$$\Big(\boldsymbol{R} + \frac{\tau}{2}\boldsymbol{A}\Big)\hat{\boldsymbol{u}}^{(m+1)} = \Big(\boldsymbol{R} - \frac{\tau}{2}\boldsymbol{A}\Big)\boldsymbol{u}^{(m)} + \frac{\tau}{2}\big(\hat{\boldsymbol{r}}^{(m+1)} + \hat{\boldsymbol{r}}^{(m)}\big), \tag{3.31}$$

where $\hat{\boldsymbol{u}}^{(m)} = \hat{\boldsymbol{u}}(m\tau)$ and $\hat{\boldsymbol{r}}^{(m)} = \hat{\boldsymbol{u}}(m\tau)$ for $m \in \mathbf{N}_0$. The iteration (3.31) has the form (3.30) if $\boldsymbol{R}$ and $-\boldsymbol{A}$ in (3.30) are replaced by

$$\boldsymbol{D} \coloneqq \boldsymbol{R} + \frac{\tau}{2}\boldsymbol{A}$$

and $\boldsymbol{R} - \frac{\tau}{2}\boldsymbol{A}$, respectively. Similar to explicit methods, the scheme (3.31) can be iterated by first Cholesky factorizing the left hand side matrix $\boldsymbol{D}$, which is always symmetric and positive-definite.

The difference between a traditional FEM system and the system resulting from our parametric equation is that in the former the mass and stiffness matrices, and thus also their linear combination, have (in general)

Table 3.1: Fill-in of the permuted matrix $\boldsymbol{D}$ that appears in the implicit midpoint time integration. Shown are approximate average values for the number of nonzero entries per row. The spatial dimension is either $d = 1$ or $d = 2$, the spline degree is determined by q, and n denotes the maximum total degree of the orthogonal polynomials.

		$n = 2$		$n = 3$	
d	q	$\overline{\text{nnz}}(\boldsymbol{D})$	$\overline{\text{nnz}}(\text{chol}(\boldsymbol{D}))$	$\overline{\text{nnz}}(\boldsymbol{D})$	$\overline{\text{nnz}}(\text{chol}(\boldsymbol{D}))$
1	0	4.36	3.60	4.90	6.04
1	1	5.15	4.84	5.98	9.46
1	2	6.45	7.88	7.78	29.2
1	3	7.24	9.71	8.86	41.7
1	∞	12.5	29.7	16.1	130.1
2	0	7.92	25.8	8.44	127.0
2	1	10.0	174.0	11.5	2011.1
2	∞	23.9	688.9	31.1	6084.5

the same sparsity pattern, whereas the parametric equation results in matrices that greatly differ in terms of their sparsity structure and number of nonzero elements. In other words, the matrix $\boldsymbol{D}$ may have more nonzero elements than the mass matrix $\boldsymbol{R}$. When using the implicit method (3.31) with Cholesky factorization, reordering the elements of $\boldsymbol{D}$ becomes essential even in one spatial dimension. Actually, in one dimension, changing the order in the Kronecker products (3.6) and (3.8) results in a block-tridiagonal system, which can be efficiently solved [11, Sec. 4.5]. However, this kind of manual reordering is not available in higher spatial dimensions. Of course, the system (3.31) can be solved by using an iterative method at every time step, but this introduces an additional numerical error which may be difficult to control.

Finding an optimal permutation such that the number of nonzero elements in the Cholesky factor is as small as possible is an NP-complete problem [24]. Thus, in practical situations one has to resort to approximations. Here, we experiment with the symmetric approximate minimum degree algorithm, which is implemented as a function `symamd` in Matlab. Table 3.1 lists some values for the number of nonzero elements (see definition (3.27)) in the Cholesky factor of the permuted matrix $\boldsymbol{D}$. The values corresponding to $d = 1$ spatial dimension are computed by assuming pure Neumann boundary conditions, uniform FEM mesh with basis functions $\{\phi_i\}_{i=1}^{25}$, and uniform

B-splines $\{\psi_p\}_{p=1}^9$ of degree $0 \leq q \leq 3$. We denote by $q = \infty$ any set of non-local diffusivity functions such as the Karhunen–Loève eigenfunctions. The actual functions are irrelevant when considering the sparsity or fill-in. Since the number of elements is $m = 24$ and we have $P = 9$ splines, the condition (3.19) holds for $q = 1, 3$ but not for $q = 0, 2$. Multivariate orthogonal polynomials with total degree $n = 2$ result in $MN = 1375$ and with $n = 3$ we obtain $MN = 5500$. We see that the number of nonzero elements in the Cholesky factor remains relatively small if splines of degree $q = 0$ are used to represent the diffusivity, but the fill-in greatly increases when the splines are supported on a larger interval. In addition, the polynomial degree n has a large impact on the fill-in. The fill-in is also affected by the parameters M and P, but this is not shown in the table 3.1. We emphasize that the used permutations are not optimal and better algorithms may result in Cholesky factors with less nonzero elements.

The two-dimensional examples in table 3.1 are performed by using the discretizations shown in figure 3.1, where $M = 25^2$ and $P = 4^2$. Again, $q = 0$ and $q = 1$ correspond to piecewise constant and piecewise linear diffusivity functions as in figure 3.1, whereas $q = \infty$ denotes non-local diffusivity functions. Multivariate orthogonal polynomials with $n = 2$ yields $MN = 95625$ and $n = 3$ results in $MN = 605625$ as in the previous section. Both the polynomial degree and the supports of the splines greatly affect the fill-in of the matrix $\boldsymbol{D}$. As a comparison, in the two-dimensional case we have $\overline{\mathrm{nnz}}(\boldsymbol{R}) \approx 6.68$ and $\overline{\mathrm{nnz}}(\mathrm{chol}(\boldsymbol{R})) \approx 14.1$, if the symmetric approximate minimum degree algorithm is applied for the mass matrix. Notice that the sparsity structure of the matrix $\boldsymbol{D}$ differs from that of the stiffness matrix $\boldsymbol{A}$, because in the latter some gradient inner products vanish non-trivially due to the chosen triangulation.

As a side note, we remark that the matrix-vector multiplications appearing in both explicit and implicit methods can be computed without ever forming the matrices $\boldsymbol{A}$ and $\boldsymbol{R} - \frac{\tau}{2}\boldsymbol{A}$ as a sum of Kronecker products. For example, a product $(\boldsymbol{X} \otimes \boldsymbol{Y})\boldsymbol{z}$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are matrices of arbitrary size, can be computed by vectorizing the product $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{X}$, where $\boldsymbol{Z}$ is a suitable matricization of the vector $\boldsymbol{z}$ [21]. If the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ are full, the computational benefits may be significant in terms of memory requirement and the number of floating point operations. For very sparse matrices, however, this kind of speed-up is usually not available, as discussed in [3]. Of course, all operations with the mass matrix $\boldsymbol{R}$ with constant diagonal blocks can be carried out block-wise, without forming any large matrix.

More advanced methods for temporal discretization may provide some benefits compared to those basic methods considered here. As an example, a first-order unconditionally stable semi-implicit method, which exploits

the properties of the discrete Galerkin system resulting from a stochastic diffusion equation, was proposed in [23]. However, the most efficient discretization strategies probably require considering the spatial and temporal dimensions together, and a simple piecewise linear FEM discretization may not be sufficient.

Chapter 4

Inverse diffusivity problem

Once we have solved the linear system (3.28) stemming from the parametric diffusion equation, we essentially have a numerical solution to the diffusion equation (2.1)–(2.2) for all spatial points $\boldsymbol{x} \in \Omega$ and for all diffusivity functions a that can be represented by the chosen splines with some coefficient vector $\boldsymbol{\vartheta} \in \Theta$. Evaluating the approximate solution for a given set of variables is usually a straightforward and computationally not very expensive task. To be precise, we have not defined how to interpolate the solution between subsequent time steps, if a finite difference method for the equation (3.28) is used. In what follows, however, we assume that the temporal discretization is fine enough so that the interpolation is not an issue.

For a single and known diffusivity a, the parametric solution is rarely of any interest, because compared to a regular non-parametric problem, solving the parametric equation requires substantially more work and also introduces additional discretization errors. However, if the diffusion coefficient is not known a priori, the parametric problem can be solved in advance, after which computing the values of u for a given diffusivity requires nothing but evaluating the parametric solution. In particular, if function values corresponding to several different diffusivities are required, for example during an iterative optimization procedure, solving one parametric equation and repeatedly substituting the appropriate parameter values can be an efficient strategy. This is exactly the topic of the present chapter. The *inverse diffusivity problem* asks to find a diffusivity function a that yields given values of the function u when all other parts of the diffusion equation (2.1)–(2.2) are known.

4.1 Estimating parameters from partial data

To begin with, we briefly discuss how the numerical solution

$$u_{M,N}(\boldsymbol{x}, t, \boldsymbol{\vartheta}) = \sum_{i=1}^{M} \sum_{j=1}^{N} \hat{u}_{i,j}(t)\phi_i(\boldsymbol{x})\varphi_j(\boldsymbol{\vartheta})$$

to the parametric diffusion equation can be easily yet quite efficiently evaluated. Thus, assume that we have obtained coefficient vectors $\hat{\boldsymbol{u}}(t) \in \mathbf{R}^{MN}$ from equation (3.28) for some values of t and a P-dimensional parameter vector $\boldsymbol{\vartheta} \in \Theta$ is given. Usually, the number of different univariate functions $\bar{\varphi}_r$, as well as the number of parameters, are relatively small and thus computing the values $\bar{\varphi}_r(\vartheta_p)$ for $0 \leq r \leq n$ and $1 \leq p \leq P$ does not require much effort. The next task is to distribute these values into a matrix which has the same structure as the $N \times P$ index matrix $\boldsymbol{\Lambda}$. The rows of the resulting matrix are then "collapsed" by computing products, whereupon the values $\varphi_j(\boldsymbol{\vartheta})$ are formed in a vector $\boldsymbol{\varphi}(\boldsymbol{\vartheta}) \in \mathbf{R}^N$. More precisely,

$$\boldsymbol{\varphi}(\boldsymbol{\vartheta}) \coloneqq \begin{bmatrix} \prod_{p=1}^{P} \bar{\varphi}_{\Lambda_{1,p}}(\vartheta_p) \\ \vdots \\ \prod_{p=1}^{P} \bar{\varphi}_{\Lambda_{N,p}}(\vartheta_p) \end{bmatrix}.$$

If we are given physical space-time coordinates $(\boldsymbol{x}_l, t_l)$ for $1 \leq l \leq L$, the values of the numerical solution at those points can be expressed as

$$\boldsymbol{U}(\boldsymbol{\vartheta}) \coloneqq \begin{bmatrix} u_{M,N}(\boldsymbol{x}_1, t_1, \boldsymbol{\vartheta}) \\ \vdots \\ u_{M,N}(\boldsymbol{x}_L, t_L, \boldsymbol{\vartheta}) \end{bmatrix} = \boldsymbol{V}\boldsymbol{\varphi}(\boldsymbol{\vartheta}),$$

where the matrix $\boldsymbol{V} \in \mathbf{R}^{L \times N}$ is defined according to

$$V_{l,j} \coloneqq \sum_{i=1}^{M} \hat{u}_{i,j}(t_l)\phi_i(\boldsymbol{x}_l). \tag{4.1}$$

If the spatial points $\boldsymbol{x}_l$ coincide with the nodes of the piecewise linear FEM basis $\{\phi_i\}_{i=1}^{M}$, the sum (4.1) reduces to a single term.

Now suppose we have an observation

$$\boldsymbol{Z}^{\diamond} \coloneqq \begin{bmatrix} u(\boldsymbol{x}_1, t_1) \\ \vdots \\ u(\boldsymbol{x}_L, t_L) \end{bmatrix}, \tag{4.2}$$

which contains values of the solution to the diffusion equation (2.1)–(2.2) for some diffusion coefficient a. The inverse problem of determining a can now be understood as finding a parameter vector $\hat{\boldsymbol{\vartheta}} \in \Theta$ such that

$$\boldsymbol{U}(\hat{\boldsymbol{\vartheta}}) = \boldsymbol{Z}^\diamond. \tag{4.3}$$

The solution diffusivity is then defined as

$$\hat{a}(\boldsymbol{x}) \coloneqq \sum_{p=1}^{P} \hat{\vartheta}_p \psi_p(\boldsymbol{x}) \tag{4.4}$$

for the chosen spline basis $\{\psi_p\}_{p=1}^P$. Note that the values in $\boldsymbol{Z}^\diamond$ may originate from a diffusivity that cannot be expressed as (4.4). Often, the system (4.3) has more equations than unknowns (i.e., $L > P$) and the solution vector $\hat{\boldsymbol{\vartheta}}$ does not necessarily exist. Therefore, it is natural to consider the minimization probem

$$\hat{\boldsymbol{\vartheta}} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\vartheta} \in \Theta} \left\| \boldsymbol{U}(\boldsymbol{\vartheta}) - \boldsymbol{Z}^\diamond \right\|_2 \tag{4.5}$$

instead. Here, $\|\cdot\|_2$ denotes the Euclidean norm. Of course, if the minimum is not unique, the definition (4.5) is slightly vague, but in practice this is not an issue.

The continuous form of the inverse diffusivity problem concerns finding a function $\hat{a} \in L^\infty(\Omega)$ that satisfies $U(\hat{a}) = Z^\diamond$, where U and $Z^\diamond$ are continuous functions defined, for example, in the whole domain Ω for some final time, or at the boundary of Ω for some time interval. For certain assumptions, the existence and uniqueness of the solution $\hat{a}$ can be guaranteed. On the other hand, sensitivity to the data $Z^\diamond$ makes it questionable whether the solution $\hat{a}$ is obtainable in practice. For theoretical details related to the continuous problem, we refer to [13, Chap. 9], where some uniqueness and stability results can be found.

As its continuous counterpart, the discrete problem (4.5) is ill-posed in the sense that, loosely speaking, small changes in the data $\boldsymbol{Z}^\diamond$ may cause large changes in the minimizing vector $\hat{\boldsymbol{\vartheta}}$. Therefore, it is important to realize that in practical situations the observation is not exact, but contains measurement errors. More precisely, we should consider a data vector

$$\boldsymbol{Z} \coloneqq \boldsymbol{Z}^\diamond + \boldsymbol{\varsigma},$$

where $\boldsymbol{\varsigma} \in \mathbf{R}^L$ contains noise values which are unknown but whose probability distribution may be known. In addition, due to approximation errors in the numerical solution, the vector $\boldsymbol{U}(\hat{\boldsymbol{\vartheta}})$ differs from $\boldsymbol{Z}$ even if the latter contains

no noise and corresponds to a diffusivity that has the form (4.4). If $\boldsymbol{U}^\diamond(\hat{\boldsymbol{\vartheta}})$ is the exact solution to the diffusion equtation with the diffusivity $\hat{a}$, we write

$$\boldsymbol{U}(\hat{\boldsymbol{\vartheta}}) = \boldsymbol{U}^\diamond(\hat{\boldsymbol{\vartheta}}) + \boldsymbol{\varrho},$$

where $\boldsymbol{\varrho} \in \mathbf{R}^L$ contains the numerical approximation errors. More generally, the vector $\boldsymbol{\varrho}$ may contain modelling errors as well, but for simplicity we do not distinguish them from approximation errors. Even though the measurement and approximation errors are often small, due to the instability of the inverse problem they must be taken into account during inversion. Regularization and statistical methods will be discussed in the following section. Other than that, the optimization of the type (4.5) is just a nonlinear least squares problem. Even though finding a good optimization algorithm may be essential for the efficiency, we do not discuss them here but instead consider (4.5) as a "black box" problem for which several well-studied algorithms exist.

The inverse diffusivity problem, as it is formulated in this section, can be considered as a bit simplified version of the thermal tomography, which can be used to detect diffusivity fluctuations inside an object [1]. At least in theory, the power of the tomography arises from the ability to infer interior information based on boundary data only. Therefore, we shall assume that the observation points $\boldsymbol{x}_l$, for $1 \leq l \leq L$, lie on the Neumann boundary Γ_{N}. With minor modifications, it would also be possible to consider the Neumann data $a\nabla u \cdot \hat{\boldsymbol{n}}$ on the Dirichlet boundary Γ_{D}. It is tempting to argue that having measurements inside the domain Ω would make it significantly easier to extract details of the diffusion coefficient. Likewise, an interior source term f in equation (2.1) would stabilize the inverse problem as pointed out in [13, Sec. 9.7]. Unfortunately, in many practical situations only the boundary, or a part of it, is accessible.

4.2 Regularization and Bayesian inversion

Consider again the non-ideal version of the minimization problem (4.5), namely

$$\hat{\boldsymbol{\vartheta}} := \operatorname*{arg\,min}_{\boldsymbol{\vartheta} \in \Theta} \left\| \boldsymbol{U}(\boldsymbol{\vartheta}) - \boldsymbol{Z} \right\|_2 = \operatorname*{arg\,min}_{\boldsymbol{\vartheta} \in \Theta} \left\| \boldsymbol{U}^\diamond(\boldsymbol{\vartheta}) - \boldsymbol{Z}^\diamond + \boldsymbol{\varrho} - \boldsymbol{\varsigma} \right\|_2, \qquad (4.6)$$

where $\boldsymbol{\vartheta}$ is a vector of length P and the other vectors have L elements. We assume that $L > P$. The approximation and measurement errors are contained in vectors $\boldsymbol{\varrho}$ and $\boldsymbol{\varsigma}$, respectively, and their values are not known. Finding a minimizing vector $\hat{\boldsymbol{\vartheta}}$, and hence the diffusion coefficient $\hat{a}$ via (4.4), means

that the function u that solves the corresponding diffusion equation more or less coincides with the erroneous data $\boldsymbol{Z}^\diamond - \boldsymbol{\varrho} + \boldsymbol{\varsigma}$. In practice, we do not want this kind of overfitting to happen. This is especially true when the problem is not stable, since slightly different errors can produce very different outcomes. Therefore, *regularization* is applied to the problem (4.6) such that the solution vector does not vary too much between different error realizations, while still agreeing with the observation to a certain extent [7]. Having a regularized and more well-posed problem also reduces the importance of choosing a proper optimization algorithm.

If the norm $\|\boldsymbol{\varrho} - \boldsymbol{\varsigma}\|_2$ is known or can be estimated, it seems reasonable to search for a vector $\hat{\boldsymbol{\vartheta}} \in \Theta$ that satisfies

$$\|\boldsymbol{U}(\hat{\boldsymbol{\vartheta}}) - \boldsymbol{Z}\|_2 \approx \|\boldsymbol{\varrho} - \boldsymbol{\varsigma}\|_2 . \tag{4.7}$$

This is called the *Morozov discrepancy principle* [14, Sec. 2.3]. Some iterative algorithms for the optimization problem (4.6) have the property that when (4.7) is used as a stopping condition, the resulting vectors $\hat{\boldsymbol{\vartheta}}$ are meaningful and have desired properties for most error realizations [14, Sec. 2.4]. If the elements in the error vector $\boldsymbol{\varrho} - \boldsymbol{\varsigma}$ are independent zero-mean Gaussian random variables with variance $\sigma^2 > 0$, it is customary to estimate $\|\boldsymbol{\varrho} - \boldsymbol{\varsigma}\|_2 \approx \sigma\sqrt{L}$ [14, Sec. 2.3]. However, in most cases the Gaussianity can safely be assumed to hold for the measurement error $\boldsymbol{\varsigma}$ only.

Perhaps the most classical regularization technique is the *Tikhonov regularization*. In its nonlinear and general form, Tikhonov regularization considers minimizing

$$T(\boldsymbol{\vartheta}) \coloneqq \|\boldsymbol{U}(\boldsymbol{\vartheta}) - \boldsymbol{Z}\|_2^2 + \lambda^2 S(\boldsymbol{\vartheta}), \tag{4.8}$$

where $S \colon \mathbf{R}^P \to \mathbf{R}$ is a suitable nonnegative penalty function and $\lambda > 0$ is a regularization parameter. Usually, one tries to find the value of λ so that the Morozov discrepancy principle (4.7) holds approximately for the vector that minimizes (4.8). The choice of the penalty function is affected by the assumptions on the diffusivity a. For example, if the diffusivity is assumed to have some smoothness and the entries of $\boldsymbol{\vartheta}$ correspond to pointwise values of a one-dimensional diffusivity with equidistant points, it is common to choose $S(\boldsymbol{\vartheta}) = \|\boldsymbol{S}^{(2)}\boldsymbol{\vartheta}\|_2^2$, where $\boldsymbol{S}^{(2)} \in \mathbf{R}^{(P-2)\times P}$ is the discretized second-order differential operator

$$\boldsymbol{S}^{(2)} \coloneqq \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix}. \tag{4.9}$$

In this way, the minimization of the Tikhonov functional T reduces to a nonlinear least squares problem. Another commonly used penalty function is $S(\boldsymbol{\vartheta}) = \|\boldsymbol{S}^{(1)}\boldsymbol{\vartheta}\|_1$, where

$$\boldsymbol{S}^{(1)} := \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbf{R}^{(P-1)\times P} \tag{4.10}$$

is a discretized first-order differential operator in one dimension. However, this *total variation regularization* results in a more difficult optimization problem. Of course, it is possible to use $S(\boldsymbol{\vartheta}) = \|\boldsymbol{S}^{(1)}\boldsymbol{\vartheta}\|_2^2$ as well.

We have already implicitly used yet another kind of regularization, namely *regularization by discretization* [15]. The idea of this ubiquitous method is very simple. As an example, consider a situation where it is known or assumed that the diffusivity is constant or at least close to constant. Using more than one constant spline is likely to yield a non-constant result and if the number of splines is large, the resulting diffusivity may have large variation. Of course, having more splines also increases the computational burden. One way to apply regularization by discretization is to iteratively solve the inverse problem with an increasing number of parameters, until a desired discrepancy (4.7) is obtained.

Sometimes the prior assumptions on the diffusion coefficient are most naturally formulated in terms of random variables and probability distributions. In addition, the measurement errors can often be assumed to follow the rules of probability, for example if the noise is assumed to be Gaussian. It follows that the diffusivity also becomes a random quantity. Neglecting the approximation error $\boldsymbol{\varsigma}$ for a moment, the posterior distribution of the parameter vector can be obtained by using the Bayes' rule as

$$p_{\text{post}}(\boldsymbol{\vartheta} \mid \boldsymbol{Z}) = \frac{p_{\text{lik}}(\boldsymbol{Z} \mid \boldsymbol{\vartheta}) p_{\text{pr}}(\boldsymbol{\vartheta})}{p(\boldsymbol{Z})},$$

where the likelihood function p_{lik} is determined by the noise model of the measurements and the prior probability p_{pr} reflects the prior assumptions. The denominator does not play any role when the maximum a posteriori estimate

$$\hat{\boldsymbol{\vartheta}}^{(\text{MAP})} := \arg\max_{\boldsymbol{\vartheta}\in\Theta} p_{\text{post}}(\boldsymbol{\vartheta} \mid \boldsymbol{Z}) \tag{4.11}$$

is sought. If the noise model is additive and Gaussian, such that $\boldsymbol{\varsigma}$ is a vector of independent identically distributed normal random variables, and if the prior also follows a multivariate Gaussian distribution with some covariance

matrix $\boldsymbol{\Gamma} \in \mathbf{R}^{P \times P}$, then maximizing the posterior in (4.11) is equivalent to minimizing some Tikhonov functional (4.8) of the least squares form.

Another Bayesian estimate is the conditional mean

$$\hat{\boldsymbol{\vartheta}}^{(\mathrm{CM})} := \int_{\Theta} \boldsymbol{\vartheta} p_{\mathrm{post}}(\boldsymbol{\vartheta} \mid \boldsymbol{Z}) \, \mathrm{d}\boldsymbol{\vartheta}.$$

If the dimension P is large, this integral can be difficult to compute. It is customary to utilize Markov chain Monte Carlo methods for approximating the posterior expectation [14, Sec. 3.6]. Such sampling methods rely on repeatedly evaluating the values of the posterior distribution. Again, it is essential that the parametric solution can be evaluated with a relatively small amount of work.

It is not unusual that the approximation errors are at least of the same order of magnitude as the measurement errors. In particular, when the forward problem is solved as a parametric differential equation, the discretization of the space $V \otimes L^2_w(\Theta)$ can produce significant errors. Thus, the Bayesian inversion paradigm may fail if those errors are not taken into account. Dealing with approximation errors is discussed in [14, Sec. 5.8] and [19]. Nevertheless, Bayesian inversion applied to parametric differential equations has been succesfully experimented with a crude posterior approximation that does not treat approximation errors explicitly [18].

Chapter 5

Numerical examples

In this chapter, we demonstrate by examples the capabilities and limitations of the numerical solution to the parametric diffusion equation. The first section considers the approximation errors resulting from the discretization of the space $V \otimes L^2_w(\Theta)$. Section 5.2 provides diffusivity reconstructions based on simulated boundary data in both one and two spatial dimensions. All computations in this chapter were run on a Linux desktop computer with 8 gigabytes of memory and Matlab 2014a.

5.1 Convergence of parametric solution

Let us illustrate how different discretization errors affect the accuracy of the parametric solution of the diffusion equation. We define a measure for the relative spatial error as

$$\widetilde{\varepsilon}_{M,N}(t) \coloneqq \frac{\|u_{M,N}(t) - u(t)\|_{L^2(\Omega)}}{\|u(t)\|_{L^2(\Omega)}},$$

where $u_{M,N}(t)$ is the solution (2.7) for some fixed parameter vector $\boldsymbol{\vartheta} \in \mathbf{R}^P$ and $u(t)$ is the exact solution of a non-parametric problem that corresponds to the diffusivity $a(\boldsymbol{x}) = a(\boldsymbol{x}, \boldsymbol{\vartheta})$. A mean error in the parameter domain is then defined by

$$\varepsilon_{M,N}(t) \coloneqq \frac{\|\widetilde{\varepsilon}_{M,N}(t)\|_{L^2_w(\Theta)}}{\|1\|_{L^2_w(\Theta)}}. \tag{5.1}$$

In practice, the integral in the numerator of (5.1) is computed by a quadrature (or cubature) rule and for each quadrature node the exact solution is computed by using standard FEM tools with a fine mesh.

In the following three examples, we study the convergence with respect to M and N, which denote the number of piecewise linear FEM basis functions ϕ_i and the number of multivariate parametric basis function φ_j, respectively. We assume that the spatial domain is the interval $\Omega = (0, \pi)$ and that the mesh $\{e_i\}_{i=1}^{M-1}$ is uniform. In each example, the diffusivity a is piecewise linear such that the derivative can be discontinuous at the midpoint $x = \pi/2$. In other words, the diffusivity consists of $P = 3$ splines and can be written as

$$a(x, \boldsymbol{\vartheta}) = \sum_{p=1}^{3} \vartheta_p \psi_p(x),$$

where the first-order splines ψ_p are the hat functions of a FEM grid with two elements. We consider only unit weight function $w = 1$, shifted Legendre polynomials and a total degree space with $N = N_{\mathrm{TD}}(3, n)$ basis functions for some n. For the quadrature rule in (5.1), we use tensorized Legendre–Gauss quadrature with 8^3 nodes [10, Sec. 3.1]. The reference solution u is computed with a uniform FEM grid with 256 elements. Both the parametric and reference solutions are integrated in time by using the implicit midpoint rule with a time step of 10^{-4}. The error resulting from the time integration is negligible in all three cases.

In the first example, we assume homogeneous Dirichlet boundary conditions, no forcing and an initial condition $u_0(x) = \sin(x)$. The mean errors $\varepsilon_{M,N}(t)$ for $M \in \{15, 31, 63\}$ and $N \in \{4, 10, 20, 35\}$ for different times $0 \leq t \leq 1$ are shown in figure 5.1. The number of parametric basis functions corresponds to (2.23) with $n \in \{1, 2, 3, 4\}$. In the left plot, the parameter domain is $\Theta = I^3$ for $I = (\frac{1}{4}, 4)$ and the right plot is obtained by choosing $I = (\frac{1}{2}, 2)$. We see that for small t, the error is dominated by the spatial discretization, but as the time increases, the discretization error of the parameter domain becomes clearly visible. This is more or less what one should expect. Indeed, if the diffusivity is constant, that is, if $\vartheta_p = a$ for some $a \in \mathbf{R}_+$ and for all $p = 1, 2, 3$, the problem admits a closed-form solution

$$u(x, t) = \sin(x) \exp(-at).$$

Now the factor $\exp(-at)$ is approximated with multivariate polynomials having a relatively small total degree. Obviously, the approximation cannot be accurate if t is large. On the other hand, if the domain Θ is small, the polynomial better fits the exponential function. This explains the difference between the left and right plots in figure 5.1.

As a second example, we consider another Dirichlet problem, this time with a homogeneous initial condition and a constant forcing term $f = 2$. Now the problem admits a non-trivial steady state solution. Figure 5.2 shows

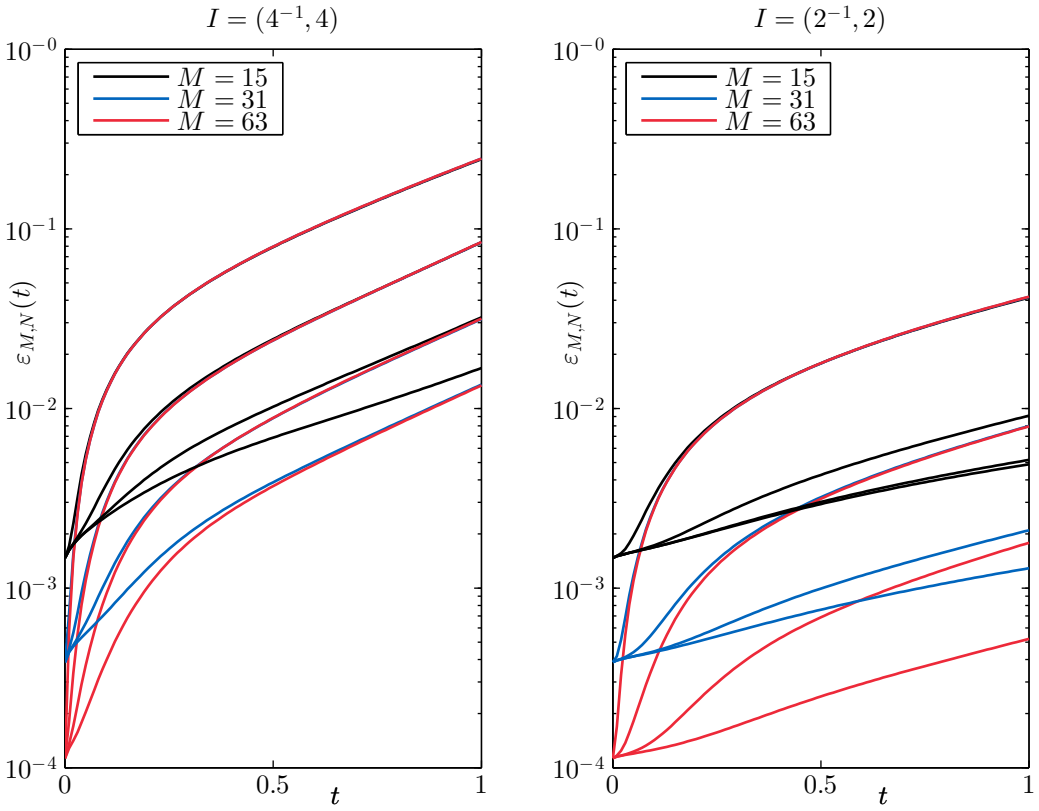

Figure 5.1: The mean error of the parametric solution $u_{M,N}(t)$ to the diffusion equation with homogeneous Dirichlet boundary conditions, the initial condition $u_0(x) = \sin(x)$ and the parameter domain $\Theta = I^3$. From top to bottom, the lines having the same color correspond to polynomial degrees $n = 1, 2, 3, 4$.

similar behavior as in the previous case. Especially with the wider parameter domain corresponding to $I = (\frac{1}{4}, 4)$, the error for a large t is dominated by the discretization of the space $L^2_w(\Theta)$. The spatial discretization error is still visible in the steady-state error, if $I = (\frac{1}{2}, 2)$.

The third example considers Neumann boundary conditions, namely the case where $g(0, t) = -20t$ and $g(\pi, t) = 20t$. These same boundary conditions will be used in the next section where we investigate the inverse problem. The initial condition is $u_0(x) = 0$ and no forcing term is present. Again, figure 5.3 shows that the dominating error results from spatial discretization if t is small and as t grows, the parametric discretization becomes more visible.

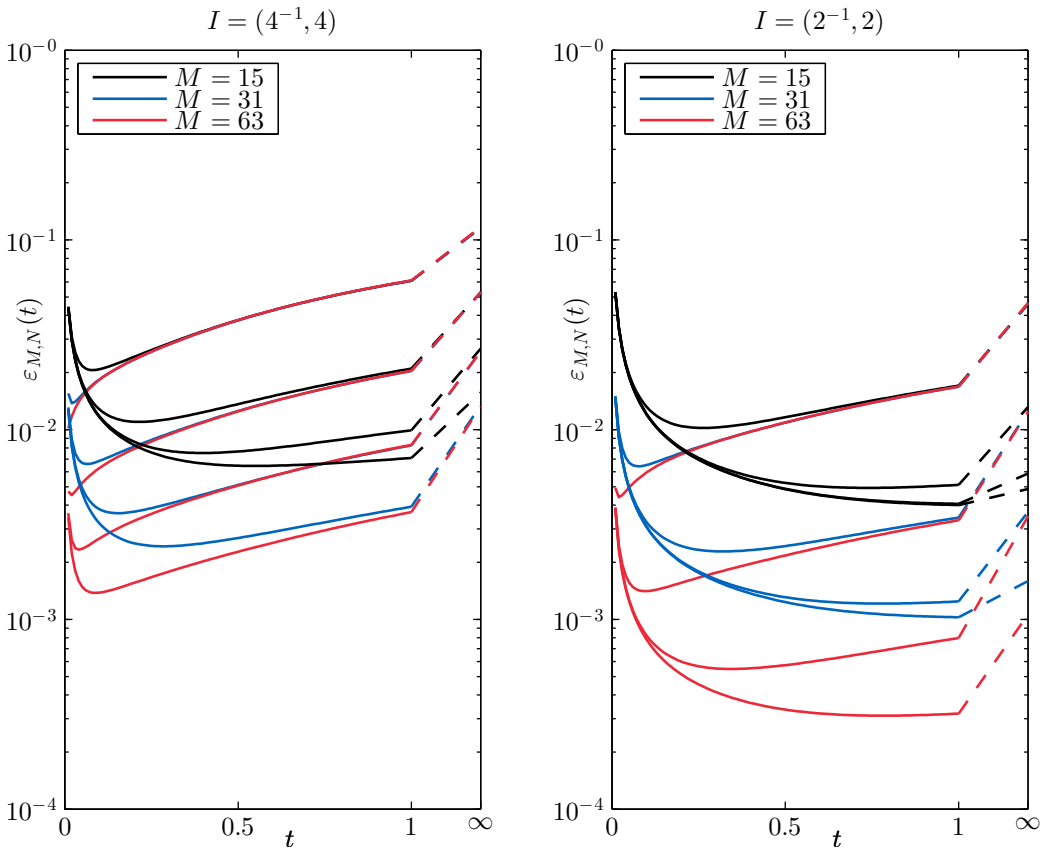

Figure 5.2: The mean error of the parametric solution $u_{M,N}(t)$ with a homogeneous intial condition, homogeneous Dirichlet boundary conditions and a constant forcing $f = 2$. The errors corresponding to $n = 1, 2, 3, 4$, as well as the steady-state error, are shown.

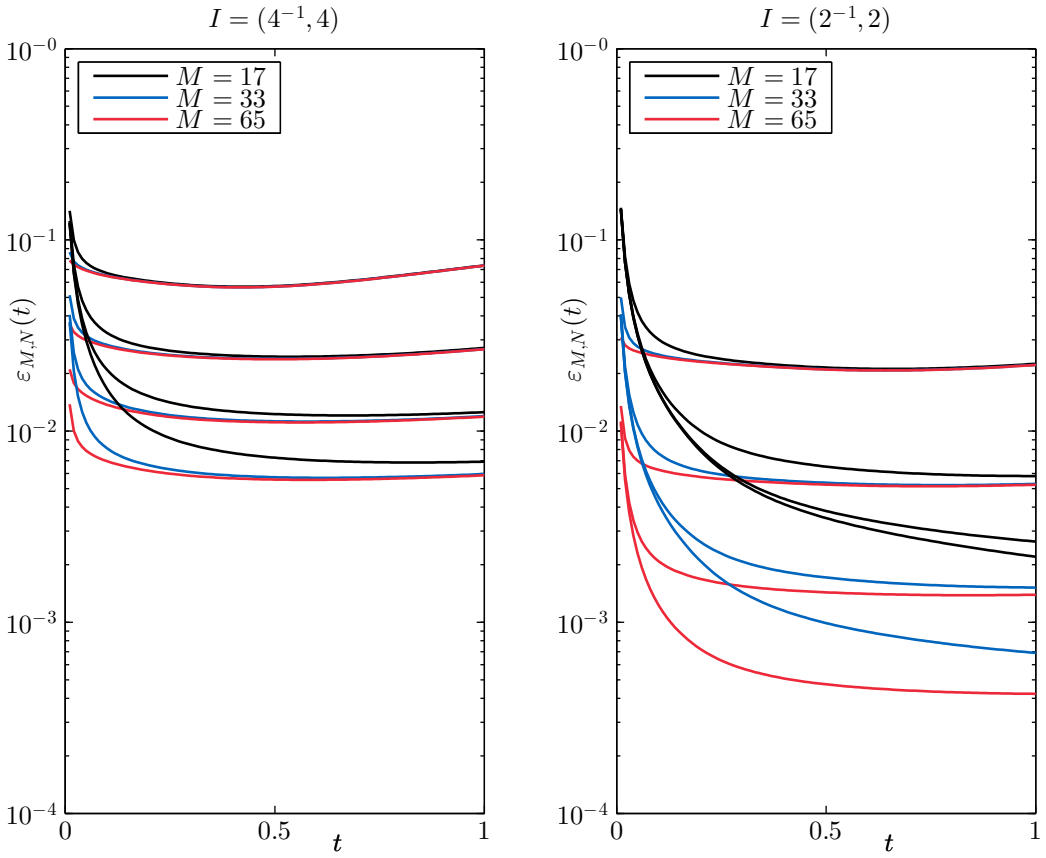

Figure 5.3: The mean error with non-trivial Neumann boundary conditions and a homogeneous initial condition. Errors corresponding to $n = 1, 2, 3, 4$ are shown.

5.2 Inverse problem and reconstructions

In this section, we illustrate how the inverse diffusivity problem can be solved with the aid of the parametric solution $u_{M,N}$. The measurement data $\boldsymbol{Z}$, possibly contaminated with artificial noise, is generated by using the finite element method with some fixed diffusivity and a fine mesh. We consider the minimization problems (4.6) and (4.8) and use the Matlab function `lsqnonlin` with its `trust-region-reflective` algorithm to perform the optimization.

The first set of experiments is conducted on the interval $\Omega = (0, 1)$. As in the third experiment of the previous section, we assume a homogeneous initial condition, no forcing, and the Neumann boundary values $g(0, t) = -20t$ and $g(1, t) = 20t$. The data $\boldsymbol{Z}$ consists of values of u at $x = 0$ and $x = 1$ for some time points. These Dirichlet boundary values are computed by using FEM with 256 linear and uniform elements and the implicit midpoint rule with a time step of 10^{-4}. For the parametric solution, multivariate Legendre polynomials with a total degree $n = 3$ in the domain $\Theta = (\frac{1}{4}, 4)^P$ are used. The piecewise linear finite element basis functions are $\{\phi_i\}_{i=1}^M$ with $M = 33$ and a uniform grid. Time integration for the equation (3.28) is performed with the implicit midpoint rule (3.31) by using a direct equation solver (i.e., Cholesky factorization) and a time step of 10^{-3}.

Let us first test four kinds of uniform B-splines, namely the transformations of those shown in figure 2.1. The exact diffusivity has the functional form $a(x) = \sin(6x) + 2.5$. The measurement times are

$$t_l \in \{0.01, 0.02, \ldots, T\} \tag{5.2}$$

with $T = 0.5$, and thus we have $L = 100$ in equation (4.2). The number of splines in each case is $P = 8$, which yields $N = N_{\mathrm{TD}}(8, 3) = 165$. We apply no regularization and the measurements contain no noise. Figure 5.4 shows that piecewise constant splines result in a relatively poor reconstruction $\hat{a}$, which is no surprise, since the sine function cannot be well approximated with only 8 pieces. Linear splines result in a much better reconstruction and the quality still improves when quadratic and cubic splines are used. Note that the underlying diffusivity is infinitely smooth, which probably favors high-order splines.

As a second example, we consider piecewise constant splines with $P = 32$ so that each spline is supported exactly on one element, or

$$\operatorname{supp}(\psi_p) = \overline{e_p}$$

for $1 \leq p \leq 32$. The diffusivity and the measurement times are the same as in the previous experiment. Without regularization, the reconstruction in the

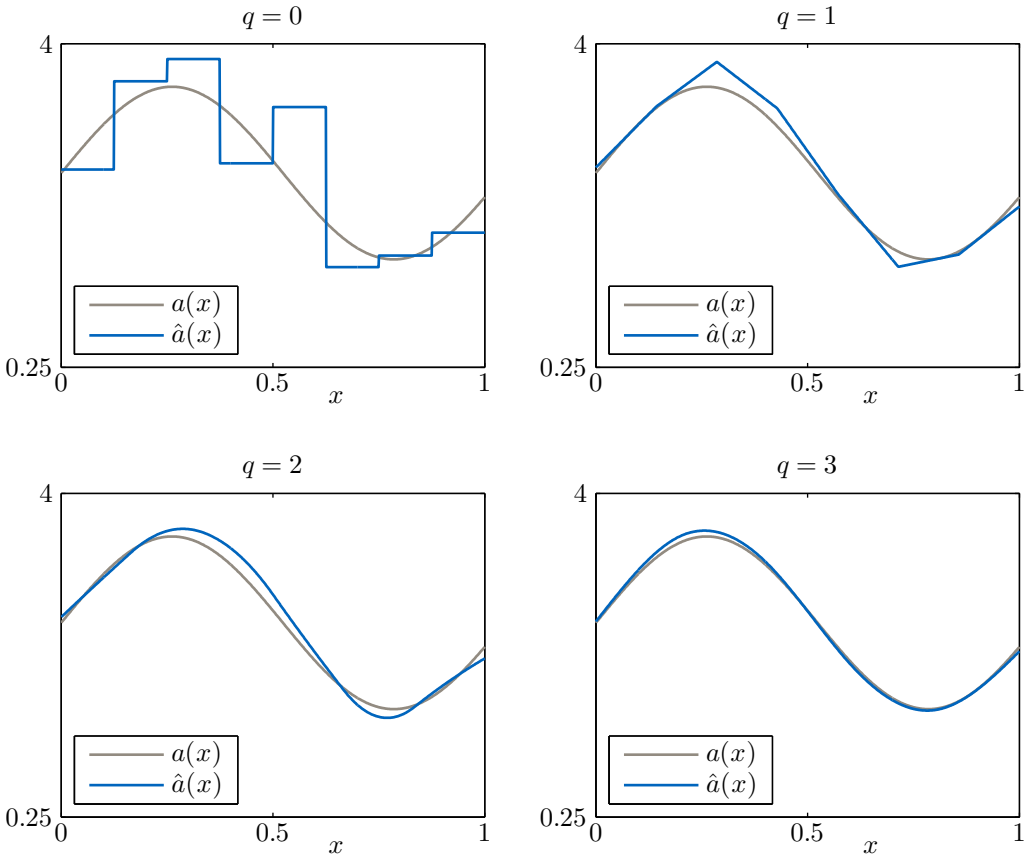

Figure 5.4: The diffusivity a and its reconstruction $\hat{a}$ as a linear combination of the qth order B-splines $\{\psi_p\}_{p=1}^{8}$.

top left plot of figure 5.5 is not better than what was already obtained with only 8 splines. However, using the smoothness matrix (4.9) and the penalty function $S(\boldsymbol{\vartheta}) = \|\boldsymbol{S}^{(2)}\boldsymbol{\vartheta}\|_2^2$ in (4.8) with $\lambda = 0.01$ or $\lambda = 0.1$ results in reconstructions that look much smoother (being, of course, still discontinuous) and match better with the target diffusivity. For $\lambda = 0, 0.01, 0.1$, the residual norms $\|\boldsymbol{U}(\hat{\boldsymbol{\vartheta}}) - \boldsymbol{Z}\|_2$ were approximately 0.003, 0.005 and 0.012, respectively. Because the measurements contain no noise, regularizing with $\lambda = 0.1$ corresponds to Morozov discrepancy principle (4.7) if the approximation error $\boldsymbol{\varrho}$ is assumed to be zero-mean Gaussian with standard deviation approximately $0.012/\sqrt{L} \approx 0.001$. Thus, it is to be expected that an additional error in the measurements does not affect the reconstruction very much if a zero-mean Gaussian noise with independent components has a standard deviation of that order of magnitude. Indeed, the bottom right plot in figure 5.5 shows that even a realization of the measurement noise with $\sigma = 0.01$ has only a

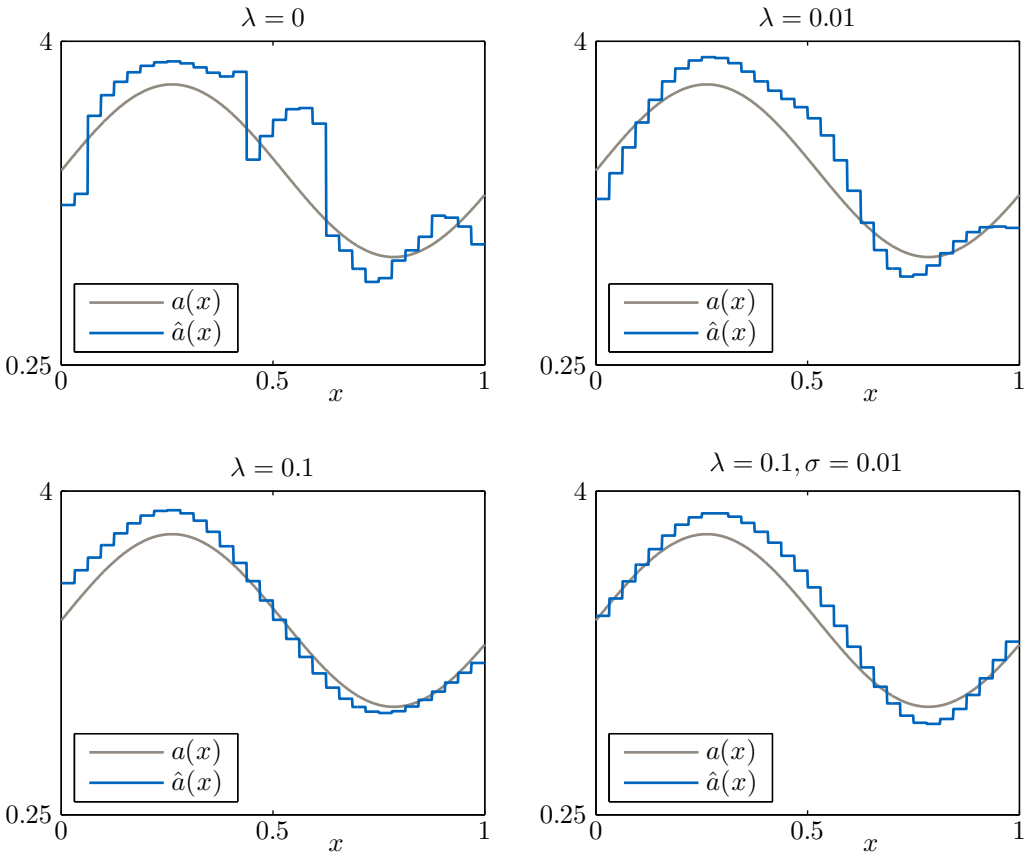

Figure 5.5: Diffusivity reconstructions by using 32 piecewise constant splines and a smoothness regularization with a regularization parameter λ. The measurement data corresponding to the last plot contains additive Gaussian noise with zero mean and standard deviation $\sigma = 0.01$.

small effect on the reconstruction.

In the third one-dimensional experiment, we consider 8 quadratic splines as in the bottom left plot of figure 5.4. This time the exact diffusivity is $a(x) = \sin(x) + 2$ and each measurement contains Gaussian noise with zero mean and standard deviation $\sigma = 0.001$. The effect of noise, which has the same realization in all four cases, is small and we apply no regularization. The top left image in figure 5.6 is obtained by using the same measurement times as earlier, that is, $T = 0.5$ in equation (5.2). For some reason, the reconstruction is worse than the one for the more rapidly oscillating sine function in figure 5.4. However, decreasing the number of time points results in better reconstructions. The other plots of figure 5.6 demonstrate that having $T = 0.4, 0.3, 0.2$ improves the quality of the reconstructions. This

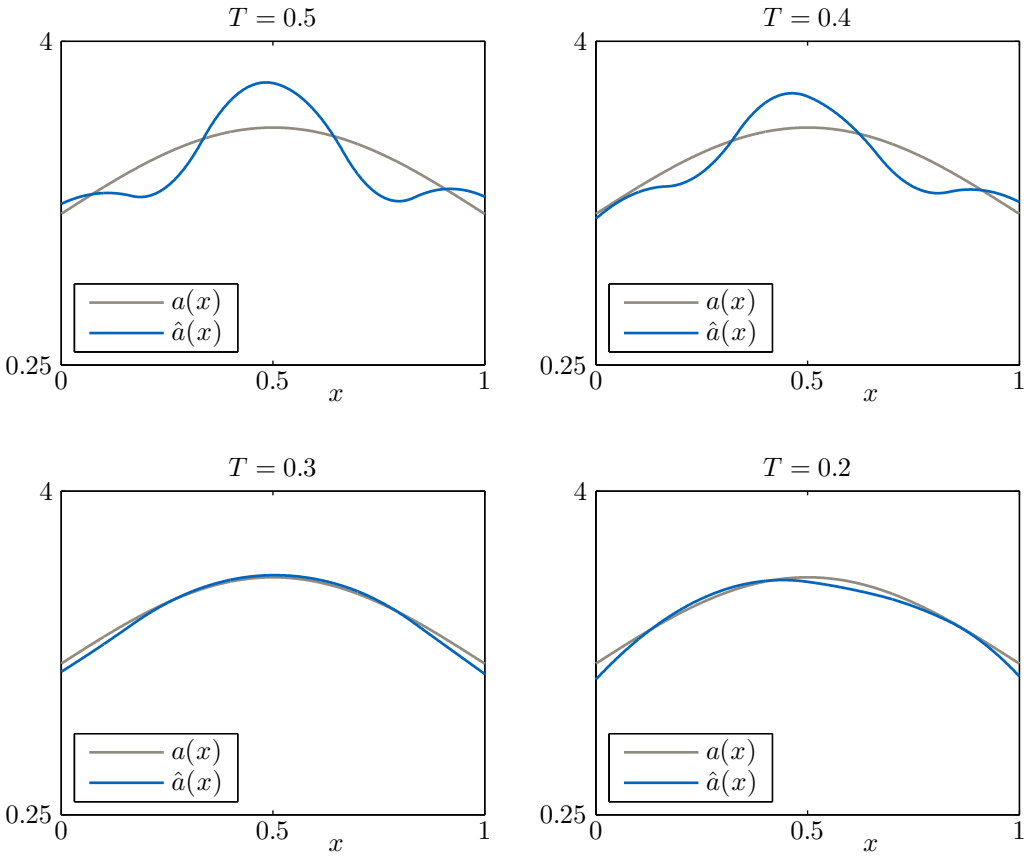

Figure 5.6: Reconstructing the diffusivity when the measurements correspond to different time intervals of the form (5.2).

phenomenon can be explained by the fact that in general the accuracy of the parametric solution becomes worse as the time increases.

To conclude, we show reconstructions in two spatial dimensions. The domain is chosen to be $\Omega = (0,1)^2$ and we assume a homogeneous initial condition with no forcing term. The Neumann boundary values are $g(\boldsymbol{x}, t) = -20t$ and $g(\boldsymbol{x}, t) = 20t$ on the boundaries $x_1 = 0$ and $x_1 = 1$, respectively, whereas the horizontal boundaries $x_2 = 0$ and $x_2 = 1$ satisfy homogeneous Neumann conditions $g = 0$. The reference solution is computed by using FEM with 32768 linear elements and the implicit midpoint rule with a time step of 10^{-4}. For the parametric solution, we choose the FEM triangulation with $M = 25^2$ basis functions ϕ_i as in figure 3.1. As in the one-dimensional case, the differential equation (3.28) is solved by using the implicit midpoint rule with a time step of 10^{-3} and the Cholesky decomposition for the matrix $\boldsymbol{D}$. Every boundary node of the coarser mesh is considered as a measurement

point and the measurement times are defined by (5.2) with some $T > 0$. We add independent zero-mean Gaussian random noise with standard deviation $\sigma = 0.001$ to each measurement. The parameter domain is discretized with multivariate Legendre polynomials of total degree $n = 2$, with the univariate polynomials being orthogonal on the interval $I = (\frac{1}{2}, 2)$.

In two dimensions, we use a first-order smoothness regularization based on the matrix $\boldsymbol{S}^{(1)}$ defined in equation (4.10). Assuming that the parameters correspond to the coefficients of the diffusivity splines $\{\psi_p\}_{p=1}^{P}$ that are equidistantly arranged column-wise or row-wise, the two-dimensional regularization matrix can be formed as

$$\boldsymbol{S} = \boldsymbol{S}^{(1)} \otimes \boldsymbol{I}_{\sqrt{P}\times\sqrt{P}} + \boldsymbol{I}_{\sqrt{P}\times\sqrt{P}} \otimes \boldsymbol{S}^{(1)}. \tag{5.3}$$

The penalty function in equation (4.8) is then chosen to be $S(\boldsymbol{\vartheta}) = \|\boldsymbol{S}\boldsymbol{\vartheta}\|_2^2$.

The first two-dimensional experiment aims to reconstruct the smooth diffusivity shown on the top left in figure 5.7. The values of the target diffusivity satisfy $0.75 \leq a(\boldsymbol{x}) \leq 1.75$ for all $\boldsymbol{x} \in \Omega$. The diffusivity representation for the parametric diffusion equation is chosen to be similar to the right image of figure 3.1, but this time the triangulation defines $P = 7^2$ basis functions ψ_p. The size of the system (3.28) then becomes $MN = 796875$ and choosing $T = 0.5$ results in a total number of $L = 4800$ observations. Without regularization, the minimization problem (4.6) results in the reconstruction shown in the top right in figure 5.7. Employing the regularization matrix (5.3) in (4.8) with a regularization parameter $\lambda = 0.1$ yields a quite accurate reconstruction, as seen in figure 5.7. An even larger regularization parameter, namely $\lambda = 1$, still results in a qualitatively correct reconstruction.

Let us then consider the target diffusivity shown in the top left image of figure 5.8. The reconstructions are based on the same parametric solution as in the previous experiment. The regularization corresponds to $\lambda = 0.1$, but now some measurements are ignored at later times. More precisely, we use the time points (5.2) with $T \in \{0.5, 0.35, 0.2\}$. The reconstructions slightly improve when less time points are used. Similar behavior was already observed in figure 5.6.

The last experiment, illustrated in figure 5.9, considers the same diffusivity as the previous experiment. This time, we use piecewise constant splines defined on a grid of size $P = 8^2$, which results in $MN = 1340625$. Despite of the large number of unknowns, performing and storing the Cholesky decomposition is still possible. The final time in (5.2) is chosen as $T = 0.35$ and the smoothness matrix (5.3) is applied to the minimization (4.8) with $\lambda = 0$ and $\lambda = 0.1$. Especially with regularization, the reconstruction is qualitatively correct. In figure 5.9, for comparison we also show a blurred version of the piecewise constant reconstruction.

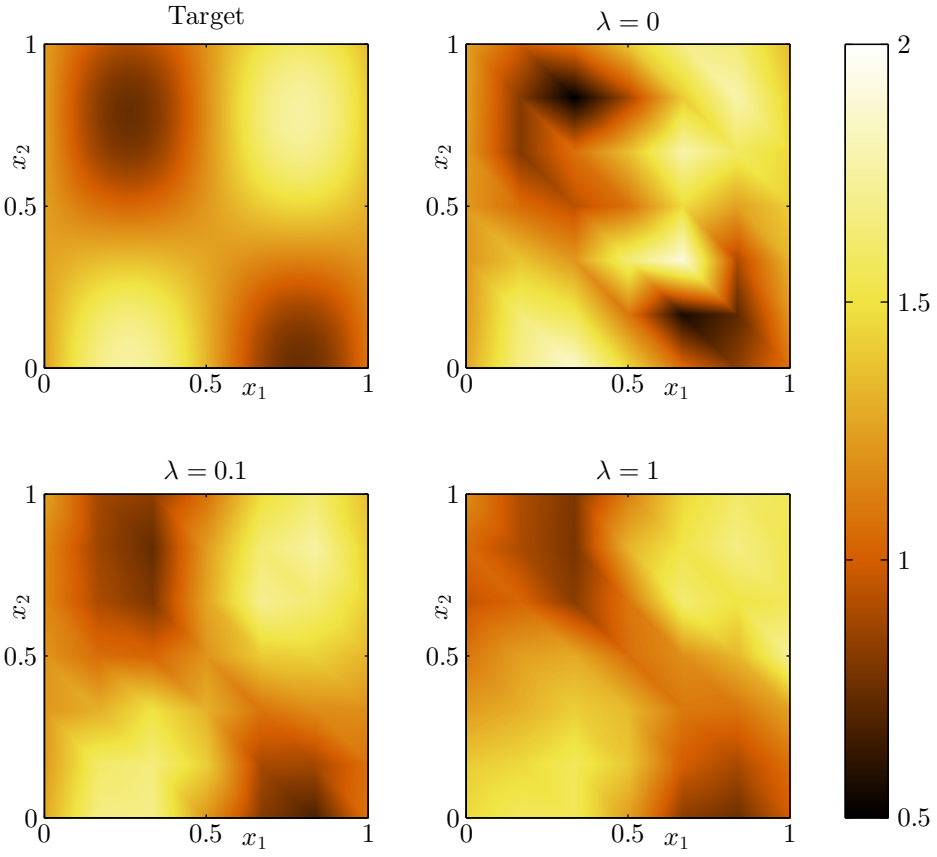

Figure 5.7: Reconstructions of a smooth diffusivity when using piecewise linear basis functions for the diffusivity and the regularization matrix (5.3) with a regularization parameter λ in (4.8).

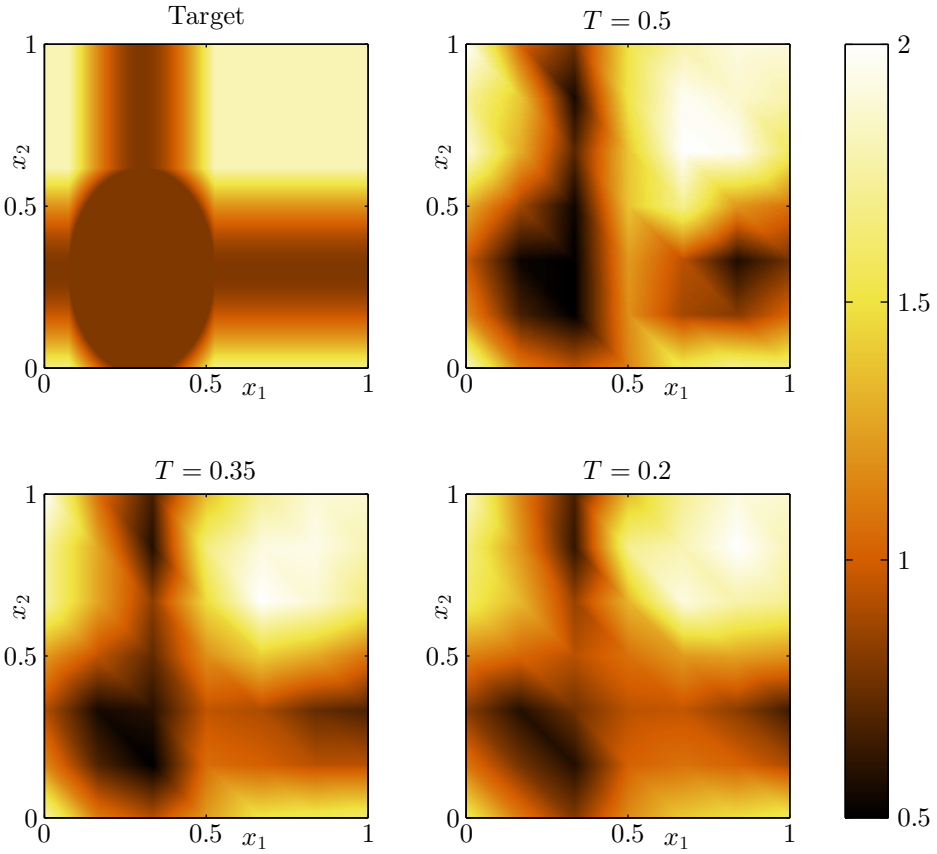

Figure 5.8: Reconstructions of a somewhat smooth diffusivity with different observation intervals, regularization parameter $\lambda = 0.1$ and the regularization matrix (5.3).

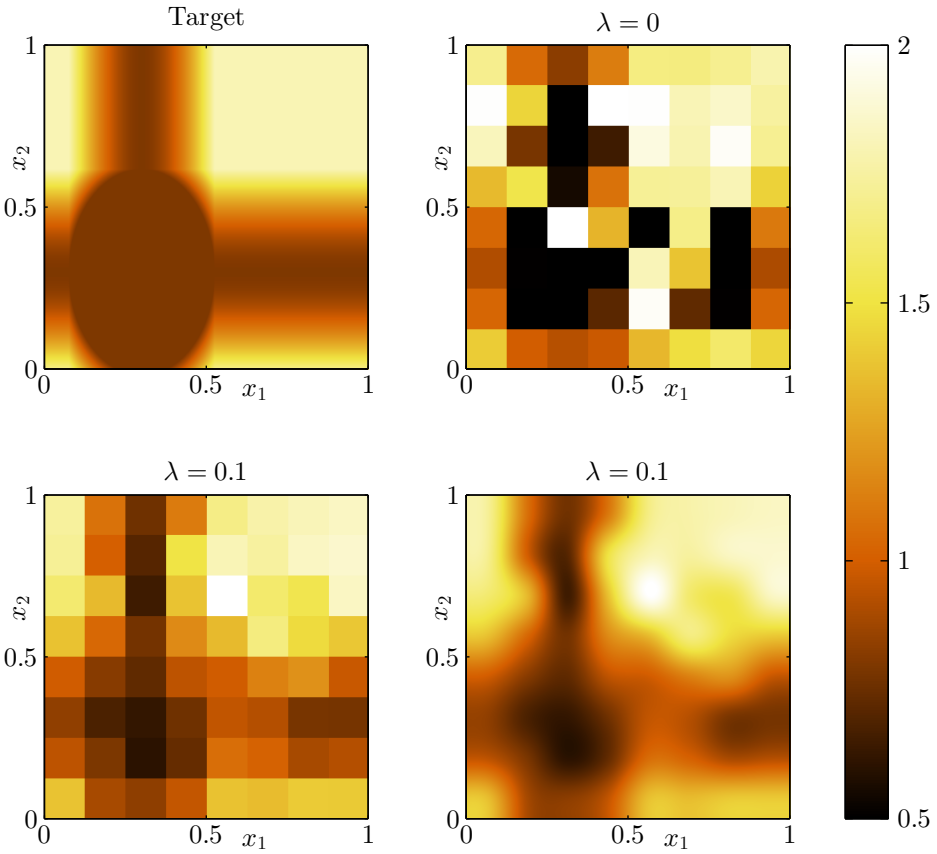

Figure 5.9: Diffusivity reconstructions with and without smoothing regularization. The bottom right image is obtained by interpolating the piecewise constant reconstruction onto a fine grid.

Chapter 6

Conclusions and future work

In this thesis, we reviewed the concept of parametric differential equations, which were subsequently utilized when an inverse boundary value problem was numerically solved. In a sense, parametric differential equations do not differ from stochastic differential equations where the function to be parametrized is treated as a random field and the parameters are random variables. However, if the parameters are not required to obey the rules of probability, the parametrization can be chosen more freely, for example to obtain computationally desirable properties.

We concentrated on spectral Galerkin methods when discretizing the parameter domain. Another option would be to use collocation methods, which result in a large number of discrete systems, each of moderate size. In contrast, Galerkin methods result in a single but very large system. The sparsity and fill-in of the Galerkin system were addressed in chapter 3, where the effect of different parametrizations was analyzed. We observed that the locality of the diffusivity basis functions had a great impact on the fill-in, if the matrices were decomposed by using Cholesky factorization. As a consequence, the memory requirements for the time-dependent diffusion equation greatly depend on the chosen parametrization and whether an explicit or implicit time integration is used. Although we mainly considered a parabolic equation, some of the observations apply to an elliptic time-independent problem as well.

Choosing efficient orthogonal bases for expressing the parameter dependence is an active research area. In this work, we mainly considered the widely used total degree polynomial spaces. Likewise, discretizing the spatial domain by using other than piecewise linear finite elements was not discussed. In addition to choosing the discretization, the spectral properties of the resulting Galerkin system require more research.

More sophisticated time discretization methods are also left for future

studies. However, the accuracy of the parametric solution is usually not limited by the time discretization. Instead, the loss of accuracy as time proceeds is caused by the need to approximate a nonlinear function with polynomials of low degree. This was demonstrated in chapter 5, where we concluded that the numerical error is dominated by the spatial discretization at early times, whereas the discretization of the parametric domain becomes the main error source later. One possible key to reducing the numerical errors is to consider an appropriately transformed equation whose solution depends more linearly on the parameters. An important generalization is to consider the case of a time-dependent diffusion coefficient.

The inverse diffusivity problem, introduced in chapter 4, was used to demonstrate the capabilities of a numerical solution to the parametric diffusion equation. The reconstructions in section 5.2 show that in simple cases the accuracy of the parametric solution is enough for obtaining qualitatively correct information on the diffusivity, even if the reconstruction is only based on noisy boundary data. Moreover, having the parametric solution at hand allows a fast evaluation and an easy use of existing minimization routines for computing the diffusivity reconstruction, even though designing the optimization algorithm for this particular task may well further improve the performance. Before the presented method is applied to any real-life problem, however, choosing a proper measurement setup requires more attention. In particular, we did not study the effect of different boundary conditions, and the optimal strategy for choosing the measurement points and times should be investigated. Nonetheless, simple Tikhonov regularization with a smoothness penalty term resulted in relatively successful reconstructions. Bayesian inversion may further provide a more flexible way to incorporate prior assumptions on the diffusivity.

References

[1] V. F. Bakirov and R. A. Kline. Diffusion-based thermal tomography. *Journal of Heat Transfer*, 127:1276–1279, 2005.

[2] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Dover Publications, second edition, 2001.

[3] P. Buchholz, G. Ciardo, S. Donatelli, and P. Kemper. Complexity of Kronecker operations on sparse matrices with applications to the solution of Markov models. Technical Report 97-66, Institute for Computer Applications in Science and Engineering (ICASE), 1997.

[4] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison. In *Spectral and High Order Methods for Partial Differential Equations*, volume 76 of *Lecture Notes in Computational Science and Engineering*, pages 43–62. Springer, 2011.

[5] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, Jr. *Spectral Methods: Fundamentals in Single Domain*. Springer, 2006.

[6] A. Chkifa, A. Cohen, R. DeVore, and Ch. Schwab. Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. *ESAIM Mathematical Modelling and Numerical Analysis*, 47:253–280, 2013.

[7] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Springer, 1996.

[8] O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann. On the convergence of generalized polynomial chaos expansions. *ESAIM Mathematical Modelling and Numerical Analysis*, 46:317–339, 2012.

[9] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 1998.

[10] W. Gautschi. *Orthogonal Polynomials: Computation and Approximation*. Oxford University Press, 2004.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, fourth edition, 2013.

[12] K. Höllig. *Finite Element Methods with B-Splines*. Society for Industrial and Applied Mathematics (SIAM), 2003.

[13] V. Isakov. *Inverse Problems for Partial Differential Equations*. Springer, second edition, 2006.

[14] J. P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer, 2005.

[15] B. Kaltenbacher and J. Offtermatt. A convergence analysis of regularization by discretization in preimage space. *Mathematics of Computation*, 81:2049–2069, 2012.

[16] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51:455–500, 2009.

[17] O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Springer, 2010.

[18] M. Leinonen, H. Hakula, and N. Hyvönen. Application of stochastic Galerkin FEM to the complete electrode model of electrical impedance tomography. *Journal of Computational Physics*, 269:181–200, 2014.

[19] A. Nissinen, L. M. Heikkinen, V. Kolehmainen, and J. P. Kaipio. Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography. *Measurement Science and Technology*, 20:105504, 2009.

[20] Ch. Schwab and C. J. Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numerica*, 20:291–467, 2011.

[21] C. F. Van Loan. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, 123:85–100, 2000.

[22] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.

[23] D. Xiu and J. Shen. Efficient stochastic Galerkin methods for random diffusion equations. *Journal of Computational Physics*, 228:266–281, 2009.

[24] M. Yannakakis. Computing the minimum fill-in is NP-complete. *SIAM Journal on Algebraic Discrete Methods*, 2:77–79, 1981.